2017

Mixture model and subgroup analysis in nationwide kidney transplant center evaluation

Lanfeng Pan
Iowa State University

Mixture model and subgroup analysis in nationwide kidney transplant center evaluation

by

Lanfeng Pan

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Yehua Li, Major Professor
Somak Dutta
Jae-Kwang Kim
Vivekananda Roy
Zhengyuan Zhu

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University
Ames, Iowa
2018

DEDICATION

I would like to dedicate this thesis to my wife Xin and my parents without whose support I would not have been able to complete this work. I would also like to thank my friends and family for their loving, guidance and financial assistance during the writing of this work.

TABLE OF CONTENTS

Page

LIST OF TABLES vii

LIST OF FIGURES ix

ACKNOWLEDGEMENTS xi

ABSTRACT xii

CHAPTER 1. LITERATURE REVIEW 1

1.1 Background 1

1.2 Generalized Linear Mixed Model 4

1.3 Finite Normal Mixture 5

1.3.1 Unbounded likelihood and unbounded Fisher information matrix . . 6

1.3.2 Identifiability issue 7

1.3.3 Consistency 9

1.3.4 Order selection via hypothesis testing 9

1.3.5 Order selection via information criterion 12

1.3.6 Order selection via penalization 13

1.4 Clustering and Subgroup Analysis 14

1.4.1 Hierarchical clustering 15

1.4.2 Convex clustering 16

1.4.3 Clustering with concave penalty 16

1.4.4 Choosing the number of clusters 18

1.5 False Discovery Rate Control 20

1.6 Penalty Functions 21

CHAPTER 2. GENERALIZED LINEAR MIXED MODELS WITH GAUSSIAN MIXTURE RANDOM EFFECTS: INFERENCE AND APPLICATION 23
2.1 Introduction 23
2.2 Model and Parameter Estimation 25
2.2.1 Model and assumptions 25
2.2.2 Model fitting 26
2.2.3 Consistency of the estimator 27
2.3 Deciding the Number of Mixture Components 28
2.3.1 Hypothesis tests on the order of the latent Gaussian mixture model 28
2.3.2 Homogeneity test 29
2.3.3 Testing for C greater than 1 31
2.4 Use False Discovery Rate Control to Classify Groups 36
2.5 Simulation Studies 38
2.5.1 Simulation 1: Estimation and random effect prediction 38
2.5.2 Simulation 2: Hypothesis tests 40
2.6 Data Analysis 44
2.6.1 Background 44
2.6.2 Model fitting 45
2.6.3 Performance evaluation 47
2.7 Summary 48
CHAPTER 3. SUBGROUP ANALYSIS AND VARIABLE SELECTION IN GENERALIZED LINEAR MIXED MODEL 50
3.1 Introduction 51
3.2 Subgroup Analysis when Transplant Center Effects Are Observed 53
3.2.1 Subgroup analysis with l_0 penalty 54
3.3 Subgroup analysis and variable selection in GLMM 57
3.3.1 Gauss-Hermite Approximation 60
3.3.2 Updating fixed effects 61
3.3.3 Clustering the random effects 63

3.3.4 Initialization and stopping . . . 65
3.4 Tuning Parameters . . . 65
3.4.1 Selecting λ_β . . . 65
3.4.2 Selecting the number of clusters . . . 66
3.5 Simulation Studies . . . 71
3.6 Data Analysis . . . 74
3.6.1 Background . . . 74
3.6.2 Model fitting . . . 75
3.7 Discussion . . . 76
CHAPTER 4. SUMMARY AND FUTURE WORK . . . 84
4.1 Summary . . . 84
4.2 Future Work . . . 85
4.2.1 Top down instead of bottom up . . . 85
4.2.2 Different penalty functions . . . 86
4.2.3 Model the fixed effects and random effects at the same time . . . 87
4.2.4 Nonparametric density based clustering . . . 88
BIBLIOGRAPHY . . . 90
APPENDIX ADDITIONAL MATERIAL FOR CHAPTER 2 . . . 98
A.1 Model fitting using EM algorithm . . . 98
A.1.1 E-Step with Gauss-Hermite quadrature approximation . . . 98
A.1.2 M-Step . . . 99
A.1.3 Stopping rule and random effect prediction . . . 100
A.2 Simulation Approach for the Asymptotic Distribution in Proposition 2.5 . . . 100
A.3 Assumptions and Consistency of the Estimator in section 2.2.2 . . . 101
A.3.1 Assumptions . . . 101
A.3.2 Proof of Proposition 2.1 . . . 103
A.4 Proof of Proposition 2.2 . . . 104
A.5 Proof of Proposition 2.3 . . . 106

A.6 Proof of Proposition 2.4 107

A.7 Proof of Proposition 2.5 108

A.8 Proof of Proposition 2.6 109

LIST OF TABLES

Page

Table 2.1 Summary for parameter estimation under Simulation Model 1 based on 200 replications. 39

Table 2.2 Summary for parameter estimation under Simulation Model 2 based on 200 replications. 39

Table 2.3 Mean squared prediction error for the random effect under Simulation Models 1 and 2. Gaussian: GLMM with Gaussian random effects; Gaussian Mixture: the proposed model; Mean: Mean Squared Prediction Error averaged over 200 replicates; Std: standard deviation of the prediction error. 41

Table 2.4 U.S. Organ Procurement and Transplantation Network data analysis: estimated fixed effect coefficients, standard errors, z-values and p-values. 46

Table 2.5 The outperforming centers detected using local false discovery rate in the kidney transplant data. 48

Table 3.1 The minimum, 1st quartile, median, mean, 3rd quartile and maximum of the Rand Index for Model 1–6 based on 200 simulations when true number of clusters is given. 73

Table 3.2 Summary of the fixed effects selected by BIC in model 0–6 based on 200 simulations. Correct selection: frequency of the correct selecting all important effects and set all redundant effects to 0; Miss nonzero covaraties: frequency of at least one important effects is not selected; Average model size: the average number of variables selected. . . . 73

Table 3.3 Number of clusters selected by each of the methods in model 0. The true number of clusters is 1. 74

Table 3.4 Number of clusters selected by each of the methods in model 1, 2 and 3 based on 200 simulations. The true number of clusters is 2. 79

Table 3.5 Number of clusters selected by each of the methods in model 4,5 and 6 based on 200 simulations. The true number of clusters is 3. . . . 80

Table 3.6 Summary of the categorical variables sex of the patient, BMI and surgery time. 81

Table 3.7 U.S. Organ Procurement and Transplantation Network data analysis: estimated fixed effect coefficients, standard errors, z-values and p-values. 82

Table 3.8 The outperforming cluster in the kidney transplant data selected by cross validation. 82

LIST OF FIGURES

Page

Figure 1.1 The number of patients N_i for each center i. 2

Figure 1.2 The death rate (x-axis) v.s. the transplant center effects (y-axis). The transplant center effects are estimated by fitting a generalized linear model by regressing the patient 5-year post-transplant survival status on the patient level demographical information as well as the transplant centers. The transplant centers are treated as a regular categorical covariate here. The point size corresponds to $\log(N_i)$ where N_i is the number of patients in transplant center i. 3

Figure 2.1 Simulation Model 1: impact of random effect assumption. Panel (a) shows results from a common generalized linear mixed model with a mis-specified Gaussian random effect assumption; Panel (b) shows results of the proposed latent Gaussian mixture model with a correctly specified number of components. In both panels, the solid curve is the true density for γ, the dashed curve is the estimated density of γ using the fitted model, and the dot-dash curve is the kernel density of the predicted random effects. 40

Figure 2.2 Empirical (dash) and asymptotic (solid) distributions of $\widetilde{T}_1$, $\widetilde{T}_2$ and $\widetilde{T}_3$ under the null hypotheses. The vertical dotted line marks the 95% quantile of the asymptotic distribution. 42

Figure 2.3 Power of the locally restricted likelihood ratio tests. Panels (a) and (c) illustrate the true density (solid) of γ under Model 3 and 4 respectively. The dashed lines represent the individual components. Panels (b) and (d) illustrate the empirical distributions (dash) of $\widetilde{T}_1$ and $\widetilde{T}_2$ comparing to the corresponding null distributions (solid). The vertical dotted line marks the 95% quantile of the null distribution. 43

Figure 2.4 (a) Estimated latent Gaussian mixture model for the kidney transplant data. The solid line and dashed line represent two components. (b) Comparison of the predicted random effects under Gaussian and Gaussian mixture model assumptions. 47

Figure 3.1 The dendrograms of Model 1–6. The x-axis is the λ_γ and the y-axis is $\widehat{\mu}_i$. Each of the lines represents how $\widehat{\mu}_i$ of transplant center i is changing along λ_γ. 78

Figure 3.2 The patient level effects $\widehat{\boldsymbol{\beta}}$ (y-axis) of the real data is very stable against the tuning parameter λ_γ (x-axis) except for some fluctations when λ_γ is very small. 81

Figure 3.3 The dendrogram of the transplant center effects (y-axis) of the real data against the tuning parameter λ_γ (x-axis). As the increase of λ_γ, the number of clusters decreases from n to 1. 83

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, Dr. Yehua Li for his insights into this area, his guidance whenever I lost direction into those technique details, his patience before I was able to overcome and his encourage when I doubt myself and my work. I would also like to thank my committee members for their efforts and contributions to this work: Dr. Somak Dutta, Dr. Jae-Kwang Kim, Dr. Vivekananda Roy and Dr. Zhengyuan Zhu.

ABSTRACT

Five year post-transplant survival rate is an important indicator on quality of care delivered by kidney transplant centers in the United States. To provide a fair assessment of each transplant center, an effect that represents the center-specific care quality, along with patient level risk factors, is often included in the risk adjustment model. In the past, the center effects have been modeled as either fixed effects or Gaussian random effects, with various merits and demerits.

We propose two new methods that allow flexible random effects distributions. The first one is a Generalized Linear Mixed Model (GLMM) with normal mixture random effects. By allowing random effects to be non homogeneous, the shrinkage effects is reduced and the predicted random effects are much closer to the truth. In addition, modeling random effects as normal mixture will essentially clustering it into different groups, which provides a natural way of evaluating the performance in the transplant center case. To decide the number of components, we do a sequential hypothesis tests.

In the second method, we propose a subgroup analysis on the random effects under the framework of GLMM. Each level of the random effect is allowed to be a cluster by itself, but clusters that are close to each other will be merged into big ones. This method provides more precise and stable estimation than fixed effects model while it has a much more flexible distributions for random effects than a GLMM with Gaussian assumption. In addition, the other effects in the model will be selected via lasso type penalty.

CHAPTER 1. LITERATURE REVIEW

1.1 Background

This research is motivated by a nationwide kidney transplant center evaluation application. The evaluation is based on the patient level health outcome, which we model with a Generalized Linear Mixed Model (GLMM). The data come from the Organ Procurement and Transplantation Network (OPTN) which helps organ transplantation institutions match waiting candidates with donated organs. The system contains many different organ transplantation databases and we only focus on kidney transplantation. The data contain nationwide kidney transplantation information from 1990 to 2008, including patient level risk factors such as demographic information, quality of the donor and which transplant center operated the surgery. The database provides the basis of a data driven evaluation of all transplant centers.

The number of patients per center is a skewed distribution with a mean of 913 and a median of 607. The smallest transplant center has only 3 patients and the largest has 5830 patients. See Figure 1.1 for a histogram and a box plot. The overall death rate of all the patients is 27.6%. The highest death rate is 61.7%, corresponding to a transplant center with 60 patients. The lowest death rate is 0%, corresponding to two small transplant centers with 3 and 4 patients respectively. As shown in Fig. 1.2 the death rate varies widely for all the 295 transplant centers. A transplant center with a higher survival rate after adjusting the patient level risk factors could mean better quality of care and would be preferred by the patients. A model that can fairly access the qualify of care delivered by the transplant centers should be of great importance to both the patients and the health policy makers.

The patient level outcome, e.g. 5-year post-transplant survival status, is modeled using a GLMM, where the random effect representing the quality of care at a transplant center and fixed effects representing the patient level risk factors mentioned above. The vast majority of the GLMM literature assumes the distribution of the random effect is Gaussian, focusing on estimating the fixed effects and treating the random effects as nuisance (Bres-

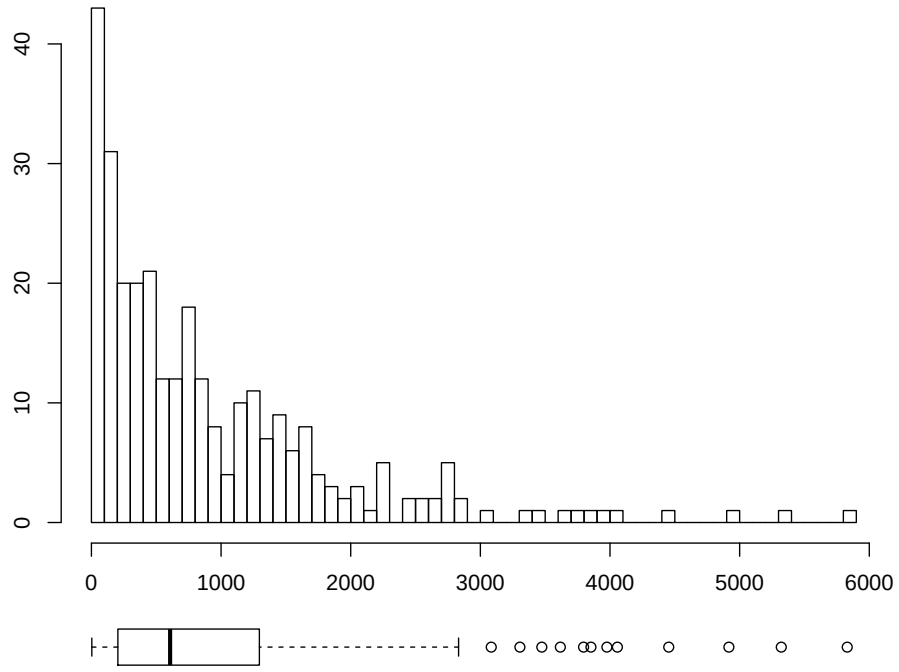


Figure 1.1 The number of patients N_i for each center i.

low and Clayton, 1993; Lin and Breslow, 1996). Even though GLMM's are typically robust against deviations from the Gaussian random effect assumption (McCulloch and Neuhaus, 2011), many authors have documented various drawbacks when the Gaussian assumption is violated, including loss of estimation efficiency (Chen et al., 2002) and reduced power for statistical tests (Litière et al., 2007). Even though the predicted random effects are relatively robust in terms of mean squared error, the distribution for the predicted random effect is highly sensitive and mostly reflects the shape of the assumed random effect distribution (McCulloch and Neuhaus, 2011). Many authors have tried to relax the Gaussian assumption and model the random effect with more flexible distributions, such as the semi-nonparametric distribution (Chen et al., 2002). Caffo et al. (2007) consider modeling the random effect with a Gaussian mixture model, but limited their investigation to binary probit models and focused on numerical performance rather than theoretical justifications.

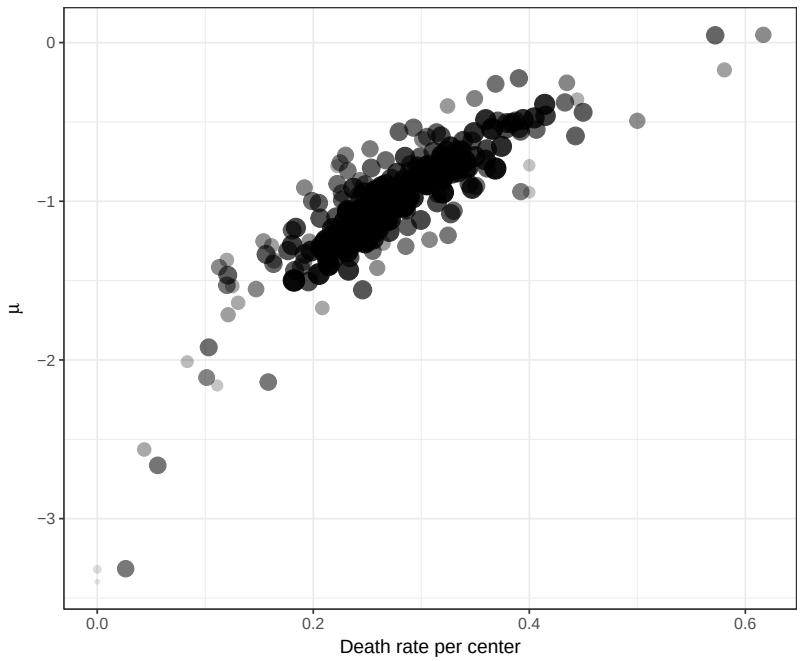


Figure 1.2 The death rate (x-axis) v.s. the transplant center effects (y-axis). The transplant center effects are estimated by fitting a generalized linear model by regressing the patient 5-year post-transplant survival status on the patient level demographical information as well as the transplant centers. The transplant centers are treated as a regular categorical covariate here. The point size corresponds to $\log(N_i)$ where N_i is the number of patients in transplant center i.

The model with a normal mixture random effects bridges the gap between fixed effects model and random effects model under normal assumption. It reduces to a model with normal random effects when the normal mixture has only one component. It becomes similar to a fixed effect model when there are same number of components as the number of transplant centers. A sequential test to decide the number of components, or the number of groups in transplant centers is proposed with theoretical justification, which can choose the correct number of clusters while controlling the first error rate.

A second approach is investigated using subgroup analysis which allows each random effect to have its own mean. Rather than choosing between modeling the transplant centers as fixed or random, this clustering approach provides a dendrogram or the clustering tree of the random effects, which shows how the model is changing from a fixed effect model to

a random effects model with fewer and fewer number of parameters. The dendrogram also gives us an insight into the structure of the heterogeneity among transplant centers besides giving the number of groups.

In Chapter 2, details about GLMM with normal mixture random effects are provided. In Chapter 3, the model using subgroup analysis is described. In Chapter 4, we discuss some possible extension and future research. In the remaining part of this chapter, we review some related topics such as GLMM, normal mixture and subgroup analysis.

1.2 Generalized Linear Mixed Model

It is common that a large set of data is collected from different sources or different strata. Some of them may have a lot of replications while the others may only have a few. One often uses random effect to represent the effect of the strata and fixed effects for other factors, i.e. a mixed model. When the response is discrete, the GLMM is further considered.

The maximum likelihood estimation of GLMM is generally difficult because the likelihood function involves an integration of a dimension equal to the number of levels of the random effect (McCulloch, 2003). The integration is even more complicated when there are multiple random effects crossing each other (Jiang, 2013). In this work, we only consider random intercept models, i.e. there is only one random effect variable and all levels of the random effect variable are independent.

Suppose there are n transplant centers and there are N_i $(i = 1, \ldots, n)$ replications in each transplant center. The overall sample size is $N = \sum_{i=1}^{n} N_i$. Each group has discrete response $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iN_i})^T$ and data matrix $\boldsymbol{X}_i$ of dimension $N_i \times J$. Denote $\boldsymbol{\beta}$ as the fixed effects and γ_i $(i = 1, \ldots, n)$ as random effects for each group. Assume the conditional distribution of Y_{ik} belongs to the canonical exponential family:

$$f(Y_{ik}|X_{ik}, \gamma_i; \boldsymbol{\beta}, \phi) = \exp\left\{\frac{Y_{ik}\xi_{ik} + b(\xi_{ik})}{a(\varphi)} + d(Y_{ik}, \xi_{ik})\right\}$$

where $a(\cdot)$, $b(\cdot)$ and $d(\cdot)$ are known functions, the canonical parameter is

$$\xi_{ik} = X_{ik}\boldsymbol{\beta} + \gamma_i \tag{1.1}$$

and $E(Y_{ik} \mid \boldsymbol{X}_{ik}, \gamma_i) = b'(\xi_{ik})$. Assume $\gamma_i \sim g(\gamma \mid \boldsymbol{\theta}_\gamma)$ then the log marginal likelihood is

$$l_n(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{X}) = \sum_{i=1}^{n} \log \int \left\{ \prod_{k=1}^{N_i} f(Y_{ik} \mid \boldsymbol{X}_{ik}, \gamma; \boldsymbol{\theta}_y) g(\gamma \mid \boldsymbol{\theta}_\gamma) \right\} d\gamma \tag{1.2}$$

where $\boldsymbol{\theta}_y = (\boldsymbol{\beta}^T, \varphi)$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_y^T, \boldsymbol{\theta}_\gamma^T)^T$.

Common estimation methods include generalized estimation equations, penalized quasi likelihood and maximum likelihood estimation with different numeric approximation. Generalized Estimation Equations is very efficient but it is more appropriate when the focus is on fixed effects (McCulloch, 2003). Its interpretation of the parameters is different from a random effects model (Lee and Nelder, 2004). Penalized quasi-likelihood and maximum likelihood with Laplace approximation suffer from biases (Breslow and Lin, 1995; Lin and Breslow, 1996). Although Breslow and Lin (1995) and Lin and Breslow (1996) give bias corrections for both models but they only work when the random effect variance is small. Monte Carlo Expectation Maximization (Booth and Hobert, 1999) approximates the integration by Monte Carlo simulations, gives consistent estimates when the Monte Carlo sample size is large enough and deals with flexible random effects distributions. But it also requires intensive computation. We choose Gauss Hermite approximation as our approach because it requires less computation than Monte Carlo Expectation Maximization and works perfectly with our normal mixtures assumptions. It also guarantees estimation consistency when enough number of nodes are used.

1.3 Finite Normal Mixture

Finite normal mixture model has a long history since Pearson (1894) and is widely used in density estimation and model based clustering. It is a very useful tool for statistical inference about heterogeneous data. Assume γ is a random variable following as normal mixture with C $(1 \leq C < \infty)$ components. Its density function is

$$g(\gamma \mid \boldsymbol{\theta}_\gamma) = \sum_{c=1}^{C} \pi_c f(\gamma \mid \mu_c, \sigma_c)$$

where f denotes the density function of $N(\mu_c, \sigma_c)$ and $\boldsymbol{\theta}_\gamma$ is the collection of components priors $(\pi_1, \ldots, \pi_C)^T$, components means $(\mu_1, \ldots, \mu_C)^T$ and component standard deviation $(\sigma_1, \ldots, \sigma_C)^T$.

Through out our research $\boldsymbol{\gamma}$ is unobserved. Instead, we observed $(y, \boldsymbol{x}^T)^T$. Thus it is more appropriate to formulate the model under a regression framework. For simplicity we assume

$$y = \boldsymbol{x}\boldsymbol{\beta} + \gamma$$

which is an analogy to (1.1). But all the following results are also applicable to the GLMM. The conditional density of y given $\boldsymbol{x}$ is

$$f(y \mid \boldsymbol{x}; \boldsymbol{\theta}) = \sum_{c=1}^{C} \pi_c f(y \mid \boldsymbol{x}\boldsymbol{\beta} + \mu_c, \sigma_c)$$

where $\boldsymbol{\theta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\theta}_\gamma^T)^T$. If we have n observations $(y_i, \boldsymbol{x}_i^T)^T$ for $(i = 1, \ldots, n)$, the log likelihood function is

$$l_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\{\sum_{c=1}^{C} \pi_c f(y_i \mid \boldsymbol{x}_i\boldsymbol{\beta} + \mu_c, \sigma_c)\} \tag{1.3}$$

and denote the maximum likelihood estimation as $\widehat{\boldsymbol{\theta}} = \{\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\pi}}^T, \widehat{\boldsymbol{\mu}}^T, \widehat{\boldsymbol{\sigma}}^T\}^T$.

Although the finite normal mixture model looks neat and simple, its theoretical property remained unknown for quite a long time partially due to the theoretical difficulties such as unbounded likelihood, unbounded Fisher information and the identifiability issue.

1.3.1 Unbounded likelihood and unbounded Fisher information matrix

The likelihood function of the normal mixture model can be infinite if any of component standard deviations is 0. To deal with this issue, Hathaway (1985) proposes a restriction on the parameters space and the restricted parameter space is

$$\Theta_{C,s} = \{\boldsymbol{\theta} \mid \sum_{c=1}^{C} \pi_c = 1, \sigma_c > 0 \text{ for } c = 1, \ldots, C, \min_{c_1, c_2}(\sigma_{c1}/\sigma_{c2}) \geq s\}$$

where s is a small constant.

Chen (2017) proposes another solution by using the following penalty function

$$-\sum_{c=1}^{C} a_n\{1/\sigma_c^2 + log(\sigma_c^2) - 1\}$$

to prevent the maximum likelihood estimator $\widehat{\sigma}_c$ from being too small. If $\widehat{\sigma}_c \to 0$ the penalty function will diverge to $-\infty$, then it essentially avoids the area with small σ_c.

The Fisher information matrix of the normal mixture model is not always finite. See the following example from Chen and Li (2009). Suppose n observations $y_1, \ldots, y_n$ are drawn from the normal mixture $(1-\pi)\ f(\gamma \mid \mu_1, \sigma_1) + \pi\ f(\gamma \mid \mu_2, \sigma_2)$. Let $\boldsymbol{x}_i = \mathbf{0}$ for $i = 1, \ldots, n$ for ease of understanding. The score function with respect to π at $\pi = 0$ is

$$\frac{\partial l_n(\boldsymbol{\theta})}{\partial \pi} \mid_{\pi=0} = \sum_{i=1}^{n} \left\{ \frac{f(y_i \mid \mu_2, \sigma_2)}{f(y_i \mid \mu_1, \sigma_1)} - 1 \right\}.$$

Then the variance of the score function is infinity if $\sigma_2^2 > 2\sigma_1^2$. We can clearly see this if we let $\mu_1 = \mu_2 = 0$ and $\sigma_1 = 1$, then the variance of the score function is

$$\frac{n}{\sqrt{2\pi}\sigma_2^2} \int \exp\{y^2(\frac{1}{2} - \frac{1}{\sigma_2^2})\} dy.$$

It is ∞ if $\sigma_2^2 > 2$. From this example, we can see the Fisher information matrix of the normal mixture model is unbounded.

1.3.2 Identifiability issue

There are three identifiability issues in the normal mixture models. The first issue is the fact that the permutation of the mixture labels does not change the density functions. This issue can be easily avoided by adding constraints such that the component means should always be increasing. If all the C components are different, we can define the constraints $\mu_1 <= \cdots <= \mu_C$. For those components with same means we further sort their components variances.

The parameter space for a model with exactly C mixture components is defined as

$$\Theta_C = \{\boldsymbol{\theta} \mid \boldsymbol{\beta} \in \mathbb{R}^p,\ \ \mu_1 < \cdots < \mu_C,\ \sum_{c=1}^{C} \pi_c = 1,\ 0 < \pi_c < 1,\ \sigma_c > 0,\ c = 1, \ldots, C\}.$$

The closure of Θ_C is

$$\bar{\Theta}_C = \{\boldsymbol{\theta} \mid \boldsymbol{\beta} \in \mathbb{R}^p,\ \sum_{c=1}^{C} \pi_c = 1,\ 0 \leq \pi_c \leq 1,\ \mu_1 \leq \cdots \leq \mu_C,\ \sigma_c \geq 0,\ c = 1, \ldots, C\},$$

which also includes the over-fitted models.

The second issue is some of the components can be identical. It means multiple possible representations of a single density function. Several normal mixtures with different number of components can be exactly equivalent when some of the components have the same mean

and variance. For example, $\tau\ N(\mu,\ \sigma^2) + (1-\tau)\ N(\mu,\ \sigma^2)$ for $0 < \tau < 1$ is equivalent to $N(\mu,\ \sigma^2)$. This is also the reason why Hathaway (1985) defined the following equivalent class before proving the consistency

$$\mathcal{F} = \left\{ \boldsymbol{\theta} \in \bar{\Theta}_C : \int_{-\infty}^{(\boldsymbol{x}', \boldsymbol{y}')} f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}) d\mu(\boldsymbol{x}, \boldsymbol{y}) = \int_{-\infty}^{(\boldsymbol{x}', \boldsymbol{y}')} f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}_0) d\mu(\boldsymbol{x}, \boldsymbol{y}) \text{ for any } (\boldsymbol{x}', \boldsymbol{y}') \right\}.$$

where $f(x, y)$ here is the joint density of $(y, \boldsymbol{x}^T)^T$ and $\bar{\Theta}_C$ is the closure of the parameter space Θ_C. If an estimation converge to $\tau\ N(\mu,\ \sigma^2) + (1-\tau)\ N(\mu,\ \sigma^2)$, it is also consistent to $N(\mu, \sigma^2)$.

The third issue is the loss of strong identifiability. In classical sense, the identifiability condition or the weak identifiability condition refers to the linear independence of the density functions in a distribution class. The strong identifiability condition (Chen, 1995, Definition 2) further imposes requirements on the first two order derivatives of the density function. A normal mixture distribution with known means is strongly identifiable (Chen, 1995). A normal mixture with known variances is also strongly identifiable. But a normal mixture with both means and variances unknown is only weakly identifiable. Denote $f(\gamma; \mu, \sigma)$ as the density of a normal distribution. It is easy to see

$$\frac{\partial^2 f(\gamma; \mu, \sigma)}{\partial \mu^2} = 2 \frac{\partial f(\gamma; \mu, \sigma)}{\partial (\sigma^2)}.$$

See Chen and Li (2009), Kasahara and Shimotsu (2015) and Ho and Nguyen (2016) for more discussions. The identifiability condition is also studied in Liu and Shao (2003) and Rousseau and Mengersen (2011).

Although consistency is obtained, the maximum likelihood estimator of the normal mixture models has a slower convergence rate than the usual $O_p(n^{-1/2})$. Chen (1995) shows the best possible convergence rate is $O(n^{-1/4})$ when the number of components is unknown but bounded. Ho and Nguyen (2016) further shows the convergence rate is $O(n^{-1/8})$ in general when the model is over fitted by one component and $O(n^{-1/12})$ when over fitted by two components. But surprisingly the convergence rate of the mixture density remains $O(n^{-1/2})$.

The $O(n^{-1/8})$ convergence rate when the model is over fitted by one component is also observed in Chen and Chen (2003) and Kasahara and Shimotsu (2015) when they develop

the asymptotic distribution of the hypothesis test statistic. The slower convergence rate also happens in our situation as reported in Chapter 2.

1.3.3 Consistency

A normal mixture density can be easily estimated via expectation-maximization algorithm and Wu (1983) shows the estimation converges to a local maximum if the order of the normal mixture is given. However the consistency is not guaranteed for a general maximum likelihood estimator and even the definition of consistency is complicated in the mixture case.

Redner (1981) proves the consistency of maximum likelihood estimation by assuming a compact parameter space containing the truth. Hathaway (1985) defines consistency in the sense of equivalent class. That means if multiple parameterization yield the same density then they are equivalent. An estimation is consistent as long as it converges to any element in the equivalent class. Hathaway (1985) further proves the consistency of maximum likelihood estimation by extending the result of Wald (1949) with an additional constraint that bounds the component variances from 0. Chen (2017) proves the consistency of penalized maximum likelihood estimation by imposing a penalty that prevents component variance from being small.

1.3.4 Order selection via hypothesis testing

Order selection is a critical issue in the normal mixture model. In our application context, the number of components means the number of groups in the transplant centers which is the primary goal of our research. One immediate choice is the likelihood ratio test. But Hartigan (1985) shows the likelihood ratio test is not able to select the true number of component due to infinite Fisher information matrix. There are several papers working on deciding the order of normal mixture via hypothesis testing. McLachlan (1987) develops a bootstrapping method to select the order of mixtures but it does not work on normal mixture due to the unbounded likelihood. Chen and Chen (2003) obtains a complicated

stochastic limiting distribution for the homogeneity test in the case of equal component variances.

Denoting C_0 as the true number of components in the model, the hypothesis testing approach seeks the answer to the question if $C_0 = C$ by observing if C components is enough to represent the data. Or formally we are testing

$$H_0 : \ C_0 = C \quad \text{versus} \quad H_1 : \ C_0 \geq C.$$

For convenience, we will name the model under null hypothesis as a reduced model and the model under alternative hypothesis as a full model. If we do sequential tests for $C = 1, 2, \ldots$ and stop when any of the tests fails to reject, we will be able to find out the true number of components subject to some chance of making mistakes.

The likelihood ratio test only works when the reduced model is an unique special case of the full model. However this is not true for the normal mixture model. Look at the following example. Let the full model be $\pi \ N(\mu_1, \sigma_1^2) + (1 - \pi) \ N(\mu_2, \sigma_2^2)$ with parameter vector $(\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)^T$ and the reduced model be $N(\mu_0, \sigma_0^2)$. The reduced model is a special case of the full model but the representation is not unique. Clearly, the full model with the following parameters $(0, \mu_1, \sigma_1^2, \mu_0, \sigma_0^2)$ or $(1, \mu_0, \sigma_0^2, \mu_2, \sigma_2^2)$ or even $(1/2, \mu_0, \sigma_0^2, \mu_0, \sigma_2^2)$ are all special cases equivalent to the reduced model. In conclusion, the limiting distribution of the log likelihood ratio is complicated when the parameters of the normal mixture model are allowed to change freely.

In light of this idea, many researchers investigate restricted likelihood ratio test. Chen and Li (2009) propose a restricted likelihood ratio test for $C_0 = 1$. They formulate the full model as

$$\tau \ N(\mu_1, \ \sigma_1^2) + (1 - \tau) \ N(\mu_2, \ \sigma_2^2)$$

where τ is a fixed constant while the other parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^T$ can be estimated via maximum likelihood estimation. The log likelihood of the full model is compared to that of the reduced model with one component. The asymptotic distribution of restricted log likelihood ratio test statistic is shown to be χ^2 distribution with degree of freedom

2. Although the full model is constrained in the restricted parameter space and is not completely maximized, our simulations show this test has very strong power.

Chen et al. (2012) further generalize the results to test arbitrary C. Their new test is formulated as $H_0 : \ C_0 = C$ versus $H_1 : \ C_0 = 2C$ for $C = 1, 2, \ldots$. They fit the reduced model first and then each component in the reduced model is split into two. As a result, the full model has $2C$ components. They show the asymptotic distribution of the test statistic is χ^2 with degree of freedom $2C$. Despite the beautiful theoretical result, it is computation intensive for large C. For each split they need to try 3 different values for τ and in combination they need to fit the full model 3^C times. In addition 100 random trials are needed each time they fit the full model. The required computation is heavy for a moderately large C.

Kasahara and Shimotsu (2015) propose to test $H_0 : \ C_0 = C$ versus $H_1 : \ C_0 = C + 1$ for $C = 1, 2, \ldots$ based on the EM test by Chen et al. (2012). The test is the same with Chen and Li (2009) and Chen et al. (2012) for $C = 1$ but only requires fitting the full model $3C$ times. Instead of splitting all components in the reduced model at the same time, they split them each at a time. The test statistic resulted from each split is asymptotically distributed as $\chi^2(2)$. The maximum of all the tests statistics gives the final restricted likelihood ratio test statistic. The test of Kasahara and Shimotsu (2015) reduces the computation but the price to pay is a complicated asymptotic distribution with no closed form. As a result, they need to generate a lot random samples from the asymptotic distribution and the p-value is estimated as the proportion of the random samples that are larger than the obtained test statistic.

The major drawback of the hypothesis testing approaches mentioned above is they all require trying many different initial values to guarantee both the restricted model and the full model are well fitted, which can be very time consuming. Besides, hypothesis test approach does not consistently select the true number of components because there is always a fixed chance of selecting the wrong model.

1.3.5 Order selection via information criterion

There are also efforts to estimate the number of components in the mixture model via information criterion. The Akaike Information Criterion (AIC) tends to overfit while some other information criterion such as integrated classification likelihood (Biernacki et al., 2000) and mixture regression criterion (Naik et al., 2007) tend to underfit (Hui et al., 2014). The Bayesian Information Criterion (BIC) is shown to be consistent for the mixture models under certain conditions (Keribin, 2000), but its power is quite weak in the normal mixture model. Xu and Chen (2015) claims the optimality of BIC for regular models cannot be extended to the non-regular normal mixture models.

Hui et al. (2014) propose AIC_{mix} which consistently selects the true order for mixture model, unlike the traditional AIC. Denote $\widehat{\boldsymbol{\theta}} = \{\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\pi}}^{\mathrm{T}}, \widehat{\boldsymbol{\mu}}^{\mathrm{T}}, (\widehat{\boldsymbol{\sigma}}^2)^{\mathrm{T}}\}^{\mathrm{T}}$ as an estimator satisfies certain conditions and $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, \ldots, \widehat{\pi}_C)^T$, $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_C)^T$, $\widehat{\boldsymbol{\sigma}} = (\widehat{\sigma}_1, \ldots, \widehat{\sigma}_C)^T$. The AIC_{mix} is defined as

$$-2\sum_{i=1}^{n}\log\left\{\sum_{c=1}^{C}\widehat{\pi}_c f(y_i|\boldsymbol{x}_i;\widehat{\mu}_c,\widehat{\sigma}_c)\right\} + 2tr\left\{\boldsymbol{I}_{comp}(\widehat{\boldsymbol{\tau}},\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\sigma}}^2)\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\sigma}}^2)\right\} + 2\sum_{c=1}^{C}\frac{\sum_{i=1}^{n}\widehat{\tau}_{ic}}{\widehat{\tau}_{ic}^2} \tag{1.4}$$

where y_i and $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ are the observed data. The function $f(y_i|\boldsymbol{x}_i;\widehat{\mu}_c,\widehat{\sigma}_c)$ is the likelihood of $N(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}} + \widehat{\mu}_c, \widehat{\sigma}_c^2)$. The posterior probability matrix $\boldsymbol{\tau}$ is consists of τ_{ic} $(i = 1, \ldots, n; c = 1, \ldots, C)$, which is the posterior probability of observation $(y_i, \boldsymbol{x}_i^T)^T$ belonging to component c. The estimation of τ_{ic} is

$$\widehat{\tau}_{ic} = \frac{\tilde{\pi}_c f(y_i|\boldsymbol{x}_i;\widehat{\boldsymbol{\beta}},\widehat{\mu}_c,\widehat{\sigma})}{\sum_{c=1}^{C}\tilde{\pi}_c f(y_i|\boldsymbol{x}_i;\widehat{\boldsymbol{\beta}},\widehat{\mu}_c,\widehat{\sigma})}.$$

Based on $\widehat{\boldsymbol{\theta}}_\gamma$, the estimation of $\boldsymbol{\tau}$ and τ_{ic} are $\widehat{\boldsymbol{\tau}}$ and $\widehat{\tau}_{ic}$ respectively. The matrix $\boldsymbol{I}(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\sigma}}^2)$ is the observed Fisher information, defined as the second derivative of

$$-\sum_{i=1}^{n}\log\left\{\sum_{c=1}^{C}\tilde{\pi}_c f(y_i|\boldsymbol{x}_i;\widehat{\mu}_c,\widehat{\sigma}_c)\right\}$$

while $\boldsymbol{I}_{comp}$ is the observed Fisher information matrix of the complete likelihood. Its estimation $\boldsymbol{I}_{comp}(\widehat{\boldsymbol{\tau}},\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\sigma}}^2)$ is the second derivative of

$$-\sum_{i=1}^{n}\sum_{c=1}^{C}\widehat{\tau}_{ic}\left\{\log(\widehat{\pi}_c) + \log f(y_i|\boldsymbol{x}_i;\widehat{\mu}_c,\widehat{\sigma}_c)\right\}.$$

Hui et al. (2014) argue the AIC_{mix} would reduce to the original AIC if all the components are far from each other and $\boldsymbol{I}_c$ and $\boldsymbol{I}$ are approximately equal. However if the components are not easy to distinguish, the penalty terms will be larger than $2C$. This can be seen in the extreme case when some components are identical. In that case $\boldsymbol{I}$ will be singular and the second term of (1.4) will be infinity. Hui et al. (2014) further prove the AIC_{mix} can consistently select the true number of components.

1.3.6 Order selection via penalization

An overfitted consistent estimator of the normal mixture model will either have diminishing π or identical components (Hui et al., 2014). This property casts light on the possibilities of selection the order via shrinkage. As Chen (1995) shows, the maximum likelihood estimation of component means converge to the truth in the order of $O(n^{-1/4})$ under the equal variance assumption. That is to say at least two component means in the model converge to the same value. Chen and Khalili (2008) propose to fit a model with C components where C is a large positive number and is believed to be larger than the true number of components. They maximize the following penalized likelihood

$$\sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \pi_c f(y_i|\boldsymbol{x}_i; \mu_c, \sigma_c) \right\} - \sum_{1 \le c_1 < c_2 \le C} p(|\mu_{c_1} - \mu_{c_2}|; \lambda) + \sum_{c=1}^{C} \lambda_\pi \log(\pi_c).$$

The penalty function $p(\cdot)$ can be some commonly used penalties such as SCAD or MCP. See section (1.6) for more details. By adjusting the penalty parameters λ, some component means will be shrunk to be exactly equal. If the right λ is used, we expect the redundant components will be forced to merge to the non-redundant ones. The third term is a penalty on $\boldsymbol{\pi}$ to prevent it from being too small. The choice of tuning parameter λ_π is not crucial and it is set to be $\log(20)$ in their simulations. They start from $\lambda = 0$ and then gradually increase the penalty parameter until a one component model is reached.

Xu and Chen (2015) propose maximizing the same penalized likelihood as Chen and Khalili (2008) but the difference is they start from one component model. By decreasing the penalty parameter, a single component will eventually be split into several. Both of the

methods are able to find a solution path by fitting the model with tuning parameters on a grid of values and then select the tuning parameter by using cross validation or BIC.

Both Chen and Khalili (2008) and Xu and Chen (2015) add an additional penalty on $\boldsymbol{\pi}$ to prevent it from being too small. Huang et al. (2017), on the contrary, aim to shrink small elements of $\boldsymbol{\pi}$ to be 0 and hence remove the redundant components. They maximize the following penalized likelihood

$$\sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \pi_c f(y_i|\boldsymbol{x}_i; \mu_c, \sigma_c) \right\} - n\lambda \sum_{c=1}^{C} D_c \{\log(\pi_c + \epsilon) - \log(\epsilon)\}$$

where D_c is the number of parameters in component c and ϵ is a small constant to bound the penalty from being infinity when $\pi_c = 0$. The small constant ϵ is set to be 10^{-6} in their simulations. A major advantage of their method is it can be applied to multivariate situations.

1.4 Clustering and Subgroup Analysis

Clustering is one of the most popular statistical topics with a long history. The normal mixture model with C ($C > 1$) components can be viewed as a model based clustering with C clusters. The same mixture distribution is assumed for all observations and each observation has a positive probability of belonging to each cluster. After obtaining the estimated mixture density the observations can then be assigned to one of the clusters by maximizing the posterior probability. But clustering is a much broader topic including many existing methods other than mixture model and its goal is much more general rather than maximizing the likelihood. Different clustering methods may achieve different goal functions. Methods such as hierarchical clustering or k-means try to minimizing the with-in cluster variations. The convex clustering defines a loss function similar to a regression problem with penalties. The clustering methods generally do not assume a common distribution for all observations. In this section, we review hierarchical clustering, subgroup analysis and methods to choose the number of clusters.

1.4.1 Hierarchical clustering

Hierarchical clustering is one of most commonly used clustering methods (Ward, 1963). It has two different implementations, agglomerative (bottom up) or divisive (top down). We describe the agglomerative algorithm here because it is closely related to our implementation used in the Chapter 3.

Denote the n observations as y_i $(i = 1, \ldots, n)$ and their dissimilarity is measured by $(y_i - y_j)^2$ for univariate case. Other dissimilarity can be defined but we focus on the simple case. Denote the clusters $\mathcal{G}_1, \ldots, \mathcal{G}_C$ as a non-overlapping split of the n observations and $|\mathcal{G}_c|$ as the group size. Given a dissimilarity $d(\cdot, \cdot)$ between clusters, the algorithm is as following.

Step 1 *Starting with* $s = 0$, $\lambda^{(0)} = 0$, $\mathcal{G}_i = \{y_i\}$ *and* $\mu_i^{(0)} = y_i$ *for* $i = 1, \ldots, C$ *where* $C = n$.

Step 2 *Update* $s \leftarrow s + 1$. *Update* $\eta_{c_1 c_2} = d(\mathcal{G}_{c_1}, \mathcal{G}_{c_2})$ *for* $1 \le c_1 < c_2 \le C$.

Step 3 *Find out the smallest* $|\eta_{c_1 c_2}|$ *and set* $\lambda^{(s)} = |\eta_{c_1 c_2}|$. *Merge* $\mathcal{G}_{c_1} \leftarrow \mathcal{G}_{c_1} \cup \mathcal{G}_{c_2}$. *Remove* $\mathcal{G}_{c_2}$ *and rename* $\mathcal{G}_{c-1} \leftarrow \mathcal{G}_c$ *for* $c_2 < c \le n$. *Update* $C \leftarrow C - 1$

Step 4 *Update* $\mu(\mathcal{G}_c) = \sum_{i \in \mathcal{G}_c} y_i / |\mathcal{G}_c|$ *and* $\mu_i^{(s)} = \mu(\mathcal{G}_c)$ *if* $i \in \mathcal{G}_c$ *for* $i = 1, \ldots, C$.

Step 5 *Go to Step 2 if* $C > 1$. *Otherwise stop and output* $\lambda^{(s)}$ *and* $\{\mu_i^{(s)}\}_{i=1}^n$ *for* $s = 0, 1 \ldots$.

There are many different definitions for the cluster dissimilarity d. Some common choices are single linkage

$$d(\mathcal{G}_{c_1}, \mathcal{G}_{c_2}) = \min_{i \in \mathcal{G}_{c_1}, j \in \mathcal{G}_{c_2}} (y_i - y_j)^2,$$

or complete linkage

$$d(\mathcal{G}_{c_1}, \mathcal{G}_{c_2}) = \max_{i \in \mathcal{G}_{c_1}, j \in \mathcal{G}_{c_2}} (y_i - y_j)^2,$$

or group average linkage

$$d(\mathcal{G}_{c_1}, \mathcal{G}_{c_2}) = \frac{\sum_{i \in \mathcal{G}_{c_1}, j \in \mathcal{G}_{c_2}} (y_i - y_j)^2}{|\mathcal{G}_{c_1}||\mathcal{G}_{c_2}|}.$$

If we define cluster mean $\mu_c = \sum_{i \in \mathcal{G}_c} y_i$, then the between group average link is $d(\mathcal{G}_{c_1}, \mathcal{G}_{c_2}) = (\mu_{c_1} - \mu_{c_2})^2$. See Murtagh (1983), Podani (1989) and Friedman et al. (2001) for more details.

1.4.2 Convex clustering

Recently there has been a lot of efforts to combine clustering together with a convex penalty, i.e. convex clustering. Given n observations y_i $(i = 1, \ldots, n)$, the goal of a convex clustering is to minimizing

$$L(\boldsymbol{\mu}) = \sum_i (y_i - \mu_i)^2 + \sum_{i_1 < i_2} w_{i_1 i_2} \lambda |\mu_{i_1} - \mu_{i2}|$$

where μ_i is the cluster mean for y_i, $w_{i_1 i_2}$ is a nonnegative weight and λ is the penalty parameter. If we view it as a regression problem, we can even more generally formulate it as the subgroup analysis

$$L(\boldsymbol{\mu}, \boldsymbol{\beta}) = \sum_i (y_i - \boldsymbol{x}_i \boldsymbol{\beta} - \mu_i)^2 + \sum_{i_1 < i_2} w_{i_1 i_2} \lambda |\mu_{i_1} - \mu_{i2}|. \tag{1.5}$$

Different penalty functions and different choices of weight $w_{i_1 i_2}$ give very different clustering results.

There are several benefits of using a convex penalty. There is a unique global minimizer to the goal function and we can find fast algorithms to solve it (Chi and Lange, 2015). It is also possible to find out the whole solution path, or the dendrogram in clustering case, in a similar way to the least angle regression (Efron et al., 2004; Radchenko and Mukherjee, 2017).

Several different implementations of convex clustering have been proposed (Hocking et al., 2011; Lindsten et al., 2011; Radchenko and Mukherjee, 2017). Among them, Chi and Lange (2015) solve the convex problem by Alternating Direction Method of Multipliers (Boyd et al., 2011, ADMM) and alternating minimization algorithm. Radchenko and Mukherjee (2017) find an elegant solution similar to the least angle regression when the penalty is l_1.

1.4.3 Clustering with concave penalty

Although convex clustering has the global minimizer it is biased (Wu et al., 2016). Several researchers' simulations show no group will be detected on the dendrogram at all using l_1 penalty (Wu et al., 2016; Ma and Huang, 2017). Several researchers try to use

concave penalties. Their goal function is formulated as

$$L(\boldsymbol{\mu}, \boldsymbol{\beta}) = \sum_i (y_i - \boldsymbol{x}_i\boldsymbol{\beta} - \mu_i)^2 + \sum_{i_1<i_2} w_{i_1 i_2} p(|\mu_{i_1} - \mu_{i2}|; \lambda). \tag{1.6}$$

Ma and Huang (2017) prefer the penalty function $p(\cdot)$ to be SCAD (Fan and Li, 2001) or MCP (Zhang, 2010) and Wu et al. (2016) use truncated lasso penalty. See section (1.6) for the definitions of these penalties. Both of them prefer $w_{i_1 i_2} = 1$, although Ma and Huang (2017) also investigate other possibilities. The dendrogram they obtained resemble those using l_1 penalties at the early stage but it gradually behave like a l_0 penalty as the size of clusters increases.

Consider the case when there are two groups, say $\mathcal{G}_1$ and $\mathcal{G}_2$ with mean μ_1 and μ_2 respectively. Assume $w_{i_1 i_2} = 1$ for now. The goal function is

$$\sum_{i\in\mathcal{G}_1} (y_i - \mu_1)^2 + \sum_{i\in\mathcal{G}_2} (y_i - \mu_2)^2 + |\mathcal{G}_1||\mathcal{G}_2| p(|\mu_1 - \mu_2|; \lambda).$$

If $\mathcal{G}_1$ and $\mathcal{G}_2$ are merged into a new group with $\mu_{new} = \frac{|\mathcal{G}_1|\mu_1 + |\mathcal{G}_2|\mu_2}{|\mathcal{G}_1| + |\mathcal{G}_2|}$, the goal function will be

$$\begin{aligned} &\sum_{i\in\mathcal{G}_1} (y_i - \mu_{new})^2 + \sum_{i\in\mathcal{G}_2} (y_i - \mu_{new})^2 \\ &\qquad = \textstyle\sum_{i\in\mathcal{G}_1} (y_i - \mu_1)^2 + \sum_{i\in\mathcal{G}_2} (y_i - \mu_2)^2 + \frac{|\mathcal{G}_1||\mathcal{G}_2||\mu_1 - \mu_2|^2}{|\mathcal{G}_1| + |\mathcal{G}_2|}. \end{aligned}$$

The difference is

$$\frac{|\mathcal{G}_1||\mathcal{G}_2||\mu_1 - \mu_2|^2}{|\mathcal{G}_1| + |\mathcal{G}_2|} - |\mathcal{G}_1||\mathcal{G}_2| p(|\mu_1 - \mu_2|; \lambda). \tag{1.7}$$

It is easy to see the penalty function is dominating as the cluster size $|\mathcal{G}_1| + |\mathcal{G}_2|$ increase, which can be as large as n. The concave penalty term will work like a hard thresholding since the first order derivative of the penalty term is exploding as $|\mathcal{G}_1| + |\mathcal{G}_2|$ increases. The difference $|\mu_1 - \mu_2|$ will be shrunk to 0 if $|\mu_1 - \mu_2| < O(\frac{1}{|\mathcal{G}_1| + |\mathcal{G}_2|})$. Otherwise the penalty term will be flat and $|\mu_1 - \mu_2|$ are not shrunk at all.

As shown in theorem 2 of Wu et al. (2016), clustering with l_0 constraint can consistently reconstruct the oracle estimator. Wu et al. (2016) claim their clustering method with truncated lasso penalty is an approximation to the clustering with l_0 penalty. Unlike convex

clustering one can only find a local minimum for concave constrained clustering. It is unclear if these clustering methods have any advantage over basic clustering methods such as hierarchical clustering. Our simulations show agglomerative hierarchical clustering yields very similar dendrogram as the algorithms in Ma and Huang (2017) and Wu et al. (2016).

1.4.4 Choosing the number of clusters

Selecting the number of clusters is a long open question in machine learning literature. Methods including Gap statistic, stability selection and information based criteria have been proposed in recent years.

Gap statistic proposed by Tibshirani et al. (2001) tries to find the largest gap between $\log(W_C)$ and its expected value under reference distribution, where W_C is the within cluster sum of squares when there are C clusters for $1 \leq C \leq C_{max}$. The maximum number of clusters C_{max} is a constant that we believe it is larger than the true number of components. The expectation and standard deviation of $\log(W_C)$, denoted as $E\{\log(W_C)\}$ and sd_C respectively, are estimated from B reference data sets drawn from a reference distribution. We run the same clustering algorithm on each reference data set to obtain multiple replications of $\log(W^*_{Cb})$. The Gap statistic is then defined as

$$Gap(C) = \frac{1}{B}\sum_{b=1}^{B}\log(W^*_{Cb}) - \log(W_C)$$

and the C clusters model is preferred than the $C+1$ clusters model if

$$Gap(C) \geq Gap(C+1) - sd_{C+1}\sqrt{1+1/B}.$$

Stability selection (Meinshausen and Bühlmann, 2010) is originally proposed for variable selection. Wang (2010) extends it to select the number of clusters by a modified cross validation with the goal of minimizing the clustering instability. Fang and Wang (2012) use the same idea but they evaluate the clustering instability by resampling instead of cross validation. Denote ψ_1 and ψ_2 as two clustering rules, which map a data point to its cluster label. The clustering distance is defined as

$$d(\psi_1, \psi_2) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[|I\{\psi_1(y_i) = \psi_1(y_j)\} - I\{\psi_2(y_i) = \psi_2(y_j)\}|]$$

where $I\{\psi_1(y_i) = \psi_1(y_j)\}$ equals 1 if $\psi_1(y_i)$ and $\psi_1(y_j)$ are equal and 0 if otherwise. Generate B pairs of independent bootstrap data $\{(y^*_{b,1}, \boldsymbol{x}^*_{b,1}), (y^*_{b,2}, \boldsymbol{x}^*_{b,2})\}$ for $b = 1, \ldots, B$. Run the clustering algorithm on both the data for each pair. We obtain clustering rules $\psi^*_{b,1,C}$ from $(y^*_{b,1}, \boldsymbol{x}^*_{b,1})$ and $\psi^*_{b,2,C}$ from $(y^*_{b,2}, \boldsymbol{x}^*_{b,2})$ for $C = 2, \ldots, C_{max}$. Then the clustering instability is defined as

$$s_B(C) = \frac{1}{B}\sum_{b=1}^{B} d(\psi^*_{b,1,C}, \psi^*_{b,2,C}).$$

The clustering distance is evaluated on the original dataset instead of the bootstrap samples. The number of clusters is chosen to be

$$\widehat{C} = \arg\min_{2 \le C \le C_{max}} s_B(C).$$

Since the clustering instability is always 0 when $C = 1$, stability selection can only choose C for $C \le 2$.

Information criteria such as AIC_{mix} or BIC are mainly used in model-based clustering. Ma and Huang (2017) choose the tuning parameter using modified BIC (Wang et al., 2009). The modified BIC is for models with diverging number of parameters. Although the number of potential parameters may diverge with n but the estimation consistency of problem (1.6) is unknown. Thus the performance of modified BIC is not satisfactory.

Consider the following example. Assume univariate observations y_i $(i = 1, \ldots, n)$ come from a standard normal distribution and $\boldsymbol{x}_i = \boldsymbol{0}$ for $i = 1, \ldots, n$. The true number of clusters should be 1 with cluster mean $\widehat{\mu}_1 = \sum_{j=1}^{n} y_i/n$ and variance $\widehat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \widehat{\mu}_1)^2/n$. The estimated log likelihood for one cluster model is

$$l_1(\widehat{\mu}_1) = \frac{n}{2} - \frac{n}{2}\log\left\{\widehat{\sigma}^2\right\}.$$

Assume a two clusters model is used anyway and two clusters $\mathcal{G}_1$ and $\mathcal{G}_2$ are obtained. Then the log likelihood

$$l_2(\widehat{\mu}_{21}, \widehat{\mu}_{22}) = \frac{n}{2} - \frac{n}{2}\log\left\{\frac{\sum_{i\in\mathcal{G}_1}(y_i - \widehat{\mu}_{21})^2 + \sum_{i\in\mathcal{G}_2}(y_i - \widehat{\mu}_{22})^2}{n}\right\}$$

is maximized when

- $\widehat{\mu}_{21} = \sum_{i\in\mathcal{G}_1} y_i/|\mathcal{G}_1|$, $\widehat{\mu}_{22} = \sum_{i\in\mathcal{G}_2} y_i/|\mathcal{G}_2|$,

- $\mathcal{G}_1$ and $\mathcal{G}_2$ evenly split the n observations, i.e. $|\mathcal{G}_1| \approx |\mathcal{G}_2| \approx n/2$ and
- $y_i > y_j$ for any $i \in \mathcal{G}_1$ and $j \in \mathcal{G}_2$ or the other way around.

The log likelihood ratio between these two models is

$$l_2(\widehat{\mu}_{21}, \widehat{\mu}_{22}) - l_1(\widehat{\mu}_1) = -\frac{n}{2}\log\left\{1 - \frac{(\widehat{\mu}_{21} - \widehat{\mu}_{22})^2|\mathcal{G}_1||\mathcal{G}_2|/n^2}{\widehat{\sigma}^2}\right\},$$

because

$$\sum_{i\in\mathcal{G}_1}(y_i - \widehat{\mu}_{21})^2 + \sum_{i\in\mathcal{G}_2}(y_i - \widehat{\mu}_{22})^2 = n\widehat{\sigma}^2 - (\widehat{\mu}_{21} - \widehat{\mu}_{22})^2|\mathcal{G}_1||\mathcal{G}_2|/n.$$

Clearly, the log likelihood ratio is of order $O_p(n)$. Recall that the two clusters model is an overfitted model. As a result, the modified BIC with a penalty in the order of $O[\log(n)\log\{\log(n)\}]$ is not appropriate to select the right number of clusters.

1.5 False Discovery Rate Control

Given multiple hypothesis tests, we are interested in filtering out the outstanding ones. The false discovery rate is the expected proportion of our discoveries that are false, or incorrect rejections. There are different approaches to select the most interesting ones.

Benjamini and Hochberg (1995) propose the BH to reject the k tests with smallest p-values where $k = \max\{i : \ p_{(i)} \leq \alpha i/n\}$ and $p_{(i)}$ is the ith smallest p-value. If no such k is found then no tests are rejected. Sun and Cai (2007) show that a z-value based procedure in more efficient than that based on p-value. The local FDR, a z-values based procedures, usually fits a normal mixture model on the z-values. Then it chooses some of the components as empirical null f_0 and the remaining as f_1. The overall distributions is

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0)f_1(z).$$

The local FDR is then defined as

$$lFDR(z) = \frac{\pi_0 f_0(z)}{f_1(z)} \tag{1.8}$$

and rejects the k tests with the smallest $lFDR$ values where $k = \max\{i : \ \sum_{i=1}^{n} lFDR_{(i)} \leq \alpha i\}$ and $lFDR_{(i)}$ is the ith smallest $lFDR$. No tests are rejected if no such k is found.

The reject regions for the above procedure are the same for all. Habiger et al. (2017) argue same rejection region for all tests is not fair if their sample size for all the tests are not equal. Tests with large sample size more easily stand out, which may not be desirable. Instead, they propose a posterior based procedure. We simplify their model to be the following. Their goal is to detect interesting effects γ_i $(i = 1, \ldots, n)$. "Interesting" means nonzero or large effects. The model for each test is

$$p(\boldsymbol{y}_i|\gamma_i; N_i)$$

where $\boldsymbol{y}_i$ is the observed data with length N_i.

The z-value based procedure would estimate γ_i, its standard error and z_i. It fits a normal mixture model on z_i $(i = 1, \ldots, n)$ and estimates $lFDR_i$ by (1.8). The posterior based procedure assumes γ_i is randomly drawn from a probability mass distribution instead of assuming normal mixture on z_i. The probability mass distribution has mass π_c on point μ_c for $c = 0, 1, \ldots, C$ with $\mu_0 = 0$. Then the mixture parameters is estimated via maximum likelihood by combining all data together

$$\sum_{i=1}^{n} \log \left\{ \sum_{c=0}^{C} \pi_c p(\boldsymbol{y}_i|\gamma_i = \mu_c; N_i) \right\}.$$

The conditional local FDR, as they put it, is defined as

$$clFDR_i = p(\gamma_i = 0|\boldsymbol{y}_i) = \frac{\pi_0 p(\boldsymbol{y}_i|\gamma_i = 0; N_i)}{\sum_{c=0}^{C} \pi_c p(\boldsymbol{y}_i|\gamma_i = \mu_c; N_i)}.$$

Similar to the local FDR, then they rank the $clFDR_i$ and reject the k tests with smallest $clFDR$ where $k = \max\{i : \sum_{i=1}^{n} clFDR_{(i)} \leq \alpha i\}$. Habiger et al. (2017) argue their reject regions are different for tests with different sample size and their conditional local FDR procedure tends to find more large effects and fewer small effects than the local FDR.

1.6 Penalty Functions

We review some popular penalty functions mentioned above, including lasso, truncated lasso, SCAD and MCP in this section. The lasso penalty is a convex penalty, while truncated lasso, SCAD and MCP are folded concave penalties.

Use β to represent the unknown parameter. The lasso penalty (Tibshirani, 1996) or the l_1 penalty is

$$p(|\beta|;\lambda) = \lambda|\beta|.$$

It has constant shrinkage no matter how large β is, which can result into bias (Fan and Li, 2001).

The SCAD penalty (Fan and Li, 2001) is proposed to adjust the bias of lasso penalty. The shrinkage effect is 0 when β is large. Fan and Li (2001) show SCAD penalty satisfies the oracle property. The definition of SCAD is

$$p(|\beta|;\lambda) = \begin{cases} \lambda|\beta| & if\ |\beta| \le \lambda; \\ \frac{2a\lambda|\beta| - \beta^2 - \lambda^2}{2(a-1)} & if\ \lambda < |\beta| \le a\lambda; \\ \frac{\lambda^2(a+1)}{2} & if\ |\beta| > a\lambda, \end{cases}$$

where $a > 2$ is required. It is usually chosen to be 3.7 as recommended by Fan and Li (2001).

The MCP penalty (Zhang, 2010) is

$$p(|\beta|;\lambda) = \begin{cases} \lambda|\beta| - \frac{\beta^2}{2a} & if\ |\beta| \le a\lambda; \\ \frac{1}{2}a\lambda^2 & if\ |\beta| > a\lambda, \end{cases}$$

where $a > 1$ is required and $a = 3$ is used in Ma and Huang (2017). The MCP is similar to SCAD in shape and also satisfies the oracle property.

The truncated lasso penalty (Shen et al., 2012), aiming to be a close approximation to l_0 penalty, is defined as

$$p(|\beta|;\lambda) = \lambda \min(|\beta|, \tau),$$

where τ is another tuning parameter.

CHAPTER 2. GENERALIZED LINEAR MIXED MODELS WITH GAUSSIAN MIXTURE RANDOM EFFECTS: INFERENCE AND APPLICATION

Lanfeng Pan, Yehua Li

Department of Statistics, Iowa State University

Kevin He, Yanming Li and Yi Li

School of Public Health & Kidney Epidemiology and Cost Center, University of Michigan

Abstract

We propose a new class of generalized linear mixed models with Gaussian mixture random effects. To overcome the weak identifiability issues of the model, we fit the model using a penalized EM algorithm, and develop sequential locally restricted likelihood ratio tests to determine the number of components in the Gaussian mixture. Our work is motivated by an application to nationwide kidney transplant center evaluation in the United States. We model the patient level post-surgery outcome by a generalized linear mixed model, which takes into account both patient level risk factors and a random effect shared by patients treated by the same transplant center. The center effect represents the center-specific quality of care and is modeled by a finite Gaussian mixture model, which provides a convenient framework to study the heterogeneity among the transplant centers and controls the false discovery rate when screening for transplant centers with non-standard performance.

Key Words: Clustering; False discovery rate; Health policy; Latent variables; Locally restricted likelihood ratio test; Penalized EM algorithm.

2.1 Introduction

The vast majority of the generalized linear mixed model (GLMM) literature assumes the distribution of the random effect is Gaussian, focusing on estimating the fixed effects and

treating the random effects as nuisance (Breslow and Clayton, 1993; Lin and Breslow, 1996). Even though GLMM's are typically robust against deviations from the Gaussian random effect assumption (McCulloch and Neuhaus, 2011), many authors have documented various drawbacks when the Gaussian assumption is violated, including loss of estimation efficiency (Chen et al., 2002) and reduced power for statistical tests (Litière et al., 2007). Even though the predicted random effects are relatively robust in terms of mean squared error, the distribution for the predicted random effect is highly sensitive and mostly reflects the shape of the assumed random effect distribution (McCulloch and Neuhaus, 2011). Many authors have tried to relax the Gaussian assumption and model the random effect with more flexible distributions, such as the semi-nonparamatric distribution (Chen et al., 2002). Caffo et al. (2007) considered modeling the random effect with a Gaussian mixture model, but limited their investigation to binary probit models and focused on numerical performance rather than theoretical justifications.

Finite Gaussian mixture models (McLachlan and Peel, 2004) are intuitively appealing for modeling non-homogeneity in a population and detecting subgroup structures. There has been a recent surge in applications of Gaussian mixture models, including clustering analysis (Huang et al., 2014), false discovery rate control (Efron, 2004; Liang and Zhang, 2008) and genetic imprinting (Li et al., 2015). In spite of its usefulness, statistical inference for Gaussian mixture models is well-known to be difficult, because many regularity conditions in parametric inference are violated in these models (Hathaway, 1985; Chen, 1995; Chen and Li, 2009). There has been much recent work in hypothesis testing on the order of finite Gaussian mixture models (Chen et al., 2012; Kasahara and Shimotsu, 2015). However, none of the existing methods are directly applicable to generalized linear mixed models.

We investigate a new class of generalized linear mixed models with Gaussian mixture random effects, propose a penalized EM algorithm to fit the proposed model and develop sequential locally restricted likelihood ratio tests to decide the number of components in the mixture model. Our work is motivated by an application on kidney transplant center evaluation, using the U.S. Organ Procurement and Transplantation Network database. We model the patient level outcome, e.g. 5-year post-transplant survival status, using a GLMM,

where the random effect representing quality of care of a transplant center follows a finite Gaussian mixture distribution. We then propose an empirical Bayes approach to classify the transplant centers using the fitted Gaussian mixture model, while controlling the false discovery rate. The results may have a strong impact on health-policy making and on patients' choice of transplant centers.

The rest of the paper is organized as follows. In Section 2.2, we introduce the model, propose an EM-based estimation procedure and establish the consistency of the procedure. To decide the number of mixture components, we propose sequential locally restricted likelihood ratio tests in Section 2.3. In Section 2.4, we propose a false discovery rate control procedure to evaluate the care qualities of the transplant centers. We conduct simulations in Section 2.5 and report the analysis of the OPTN kidney transplant data in Section 2.6. Finally, we end the paper with concluding remarks in Section 2.7. Detailed algorithms, technical proofs and additional simulation results are deferred to the supplementary material.

2.2 Model and Parameter Estimation

2.2.1 Model and assumptions

Suppose that there are n independent groups, e.g. transplant centers, each with N_i subjects, which brings the total sample size to be $N = \sum_{i=1}^{n} N_i$. Let Y_{ik} be the outcome variable of the kth patient treated at the ith group and let $\boldsymbol{X}_{ik} \in \mathbb{R}^p$ be the subject level covariate, $k = 1, \ldots, N_i$, $i = 1, \ldots, n$. Denote by $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iN_i})^{\mathrm{T}}$, $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iN_i})^{\mathrm{T}}$, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^{\mathrm{T}}$ where γ_i is the random effect representing the effect of the ith group. In our motivating example, γ_i represents the quality of care delivered by the ith transplant center. The conditional density of Y_{ik}, given $\boldsymbol{X}_{ik}$ and γ_i, belongs to the canonical exponential family:

$$f(Y_{ik} \mid \boldsymbol{X}_{ik}, \gamma_i; \boldsymbol{\beta}, \varphi) = \exp\left\{ \frac{Y_{ik}\xi_{ik} + b(\xi_{ik})}{a(\varphi)} + d(Y_{ik}, \varphi) \right\}, \tag{2.1}$$

where $a(\cdot)$, $b(\cdot)$ and $d(\cdot)$ are known functions, $\xi_{ik} = \boldsymbol{X}_{ik}^T\boldsymbol{\beta} + \gamma_i$ is the canonical parameter with $\mathrm{E}(Y_{ik} \mid \boldsymbol{X}_{ik}, \gamma_i) = b'(\xi_{ik})$, and φ is a nuisance parameter. Here $\boldsymbol{X}_{ik}$ does not contain

the intercept and γ_i has a nonzero mean. We also assume that Y_{ik} and $Y_{ik'}$ are independent given γ_i, for any $k \neq k'$. In our transplant center evaluation application, we consider a binary response variable: $Y_{ik} = 1$ if the patient died within 5 years after transplant; -1 otherwise. In the dataset, there was essentially no censoring within the first 5 years since the transplant patients' survival information had been closely monitored and tracked. With that, model (2.1) becomes $f(Y_{ik} \mid \boldsymbol{X}_{ik}, \gamma_i; \boldsymbol{\beta}) = \{1 + \exp(-\xi_{ik} Y_{ik})\}^{-1}$.

Assume that the groups belong to C subpopulations and the cth subpopulation can be described by a Gaussian distribution with mean μ_c and variance σ_c^2, $c = 1, \ldots, C$. The density of γ_i is $g(\gamma \mid \boldsymbol{\theta}_\gamma) = \sum_{c=1}^{C} \pi_c f_c(\gamma \mid \mu_c, \sigma_c)$, where $f_c(\gamma \mid \mu_c, \sigma_c) = \sigma_c^{-1} \phi\{(\gamma - \mu_c)/\sigma_c\}$, $\phi(\cdot)$ is the standard Gaussian density, $\pi_c \in [0, 1]$ is the weight for subpopulation c, $\sum_{c=1}^{C} \pi_c = 1$, and $\boldsymbol{\theta}_\gamma = (\mu_1, \ldots, \mu_C, \sigma_1^2, \ldots, \sigma_C^2, \pi_1, \ldots, \pi_C)^T$ collects the parameters in $g(\gamma)$.

2.2.2 Model fitting

Though conceptually appealing, Gaussian mixture models possess some undesirable properties, such as slower convergence rate for parameter estimation when the number of components is unknown (Chen, 1995); unbounded likelihood when any of the component variance parameters σ_c^2 goes to 0 (Hathaway, 1985); and infinite Fisher information on some boundary points of the parameter space (Chen and Li, 2009). The solution to these problems in the literature is to either restrict the value of the parameters away from the boundaries (Hathaway, 1985) or include a penalty function to prevent any σ_c from converging to 0 (Chen et al., 2008; Chen and Li, 2009).

We propose to adopt the latter strategy by maximizing a penalized likelihood

$$l_{pen}(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{X}) = l_n(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{X}) + \sum_{c=1}^{C} p_n(\sigma_c^2), \tag{2.2}$$

where $\boldsymbol{Y} = (\boldsymbol{Y}_1^{\mathrm{T}}, \ldots, \boldsymbol{Y}_n^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{X} = (\boldsymbol{X}_1^{\mathrm{T}}, \ldots, \boldsymbol{X}_n^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_y^{\mathrm{T}}, \boldsymbol{\theta}_\gamma^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\theta}_y = (\boldsymbol{\beta}^{\mathrm{T}}, \varphi)^{\mathrm{T}}$, and

$$l_n(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{X}) = \sum_{i=1}^{n} \log \int \left\{ \prod_{k=1}^{N_i} f(Y_{ik} \mid \boldsymbol{X}_{ik}, \gamma; \boldsymbol{\theta}_y) g(\gamma \mid \boldsymbol{\theta}_\gamma) \right\} d\gamma. \tag{2.3}$$

In all of our numerical studies, we use the following penalty proposed by Chen and Li (2009)

$$p_n(\sigma^2; \widehat{\sigma}^2_{pilot}) = -a_n \{\widehat{\sigma}^2_{pilot}/\sigma^2 + \log(\sigma^2/\widehat{\sigma}^2_{pilot}) - 1\}, \tag{2.4}$$

where $\widehat{\sigma}^2_{pilot}$ is a pilot estimate for the variance of γ. One possible choice of $\widehat{\sigma}^2_{pilot}$ is the variance estimator assuming the γ_i are i.i.d. Gaussian variables. When $a_n = o_p(n^{1/4})$, the penalty function in (2.4) satisfies the assumptions for our asymptotic theory. A similar requirement on a_n is made by Chen et al. (2012). In all of our numerical studies, we choose a_n using the empirical formula (23) in Kasahara and Shimotsu (2015).

To facilitate an EM algorithm, define $\boldsymbol{L}_i = (L_{i1}, \ldots, L_{iC})^T \sim$ Multinomial $(\pi_1, \ldots, \pi_C)$ as a latent random vector of subpopulation memberships, where $L_{ic} = 1$ if γ_i belongs to component c and $L_{ic} = 0$ otherwise. Then the likelihood function for the complete data, comprising of both observed and latent variables, is

$$l_{comp}(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\gamma}, \boldsymbol{L}) = \sum_{i=1}^{n} \ell_{i,comp}(\boldsymbol{\theta}; \boldsymbol{Y}_i, \boldsymbol{X}_i, \gamma_i, \boldsymbol{L}_i),$$

where

$$\ell_{i,comp}(\boldsymbol{\theta}; \boldsymbol{Y}_i, \boldsymbol{X}_i, \gamma_i, \boldsymbol{L}_i) = \log f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma_i; \boldsymbol{\theta}_y) + \sum_{c=1}^{C} L_{ic}[\log \pi_c - \frac{1}{2}\log(\sigma_c^2) + \log \phi\{(\gamma_i - \mu_c)/\sigma_c\}]$$

and $f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma_i; \boldsymbol{\theta}_y) = \prod_{k=1}^{N_i} f(Y_{ik} \mid \boldsymbol{X}_{ik}, \gamma_i; \boldsymbol{\theta}_y)$. We estimated the parameters by maximizing the penalized likelihood while treating $\boldsymbol{\gamma}$ and $\boldsymbol{L}$ as missing data. The detailed algorithm is provided in the supplementary material.

2.2.3 Consistency of the estimator

The parameter space for a model with exactly C mixture components is

$$\Theta_C = \{\boldsymbol{\theta} \mid \boldsymbol{\beta} \in \mathbb{R}^p, \ \ \mu_1 < \cdots < \mu_C, \ \sum_{c=1}^{C} \pi_c = 1, \ 0 < \pi_c < 1, \ \sigma_c > 0, \ c = 1, \ldots, C\}.$$

The closure of Θ_C is

$$\bar{\Theta}_C = \{\boldsymbol{\theta} \mid \boldsymbol{\beta} \in \mathbb{R}^p, \ \sum_{c=1}^{C} \pi_c = 1, \ 0 \le \pi_c \le 1, \ \mu_1 \le \cdots \le \mu_C, \ \sigma_c \ge 0, \ c = 1, \ldots, C\},$$

which also includes the over-fitted models. In other words, $\bar{\Theta}_C$ admits models with the true number of components C_0 strictly less than C, in which case a redundant component c can be parameterized in $\bar{\Theta}_C$ in multiple ways, such as setting either $\pi_c = 0$ or $(\mu_c, \sigma_c) = (\mu_{c'}, \sigma_{c'})$ for some $c' \neq c$. Let $\boldsymbol{\theta}_0 \in \bar{\Theta}_C$ be one parameterization for the true density of γ, and

$f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta})$ be the joint distribution function of $(\boldsymbol{X}, \boldsymbol{Y})$ associated with the likelihood in (2.3). Following Hathaway (1985), define

$$\mathcal{F} = \left\{ \boldsymbol{\theta} \in \bar{\Theta}_C : \int_{-\infty}^{(\boldsymbol{x}', \boldsymbol{y}')} f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}) d\mu(\boldsymbol{x}, \boldsymbol{y}) = \int_{-\infty}^{(\boldsymbol{x}', \boldsymbol{y}')} f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}_0) d\mu(\boldsymbol{x}, \boldsymbol{y}) \text{ for any } (\boldsymbol{x}', \boldsymbol{y}') \right\}.$$

All parameters in $\mathcal{F}$ lead to the same mixture density for γ, stressing the lack of identifiability in finite Gaussian mixture models and their fundamental difference from other commonly used parametric models. Denote the maximum penalized likelihood estimator under a C-component mixture model by $\widehat{\boldsymbol{\theta}}_C = \arg\max_{\boldsymbol{\theta} \in \bar{\Theta}_C} l_{pen}(\boldsymbol{\theta})$. The consistency for $\widehat{\boldsymbol{\theta}}_C$ is established in the following proposition, the proof of which is relegated to the supplementary material.

Proposition 2.1 *Under Assumptions 1-6 in the supplementary material,* $\widehat{\boldsymbol{\theta}}_C$ *is consistent in the sense that* $\inf_{\boldsymbol{\theta}^* \in \mathcal{F}} \|\widehat{\boldsymbol{\theta}}_C - \boldsymbol{\theta}^*\| \to 0$ *in probability.*

Proposition 2.1 implies that we can estimate the mixture density consistently, but this is not necessarily true for the parameters since $\boldsymbol{\theta}_\gamma$ is not unique if we over fit the model by including more mixture components.

2.3 Deciding the Number of Mixture Components

2.3.1 Hypothesis tests on the order of the latent Gaussian mixture model

Deciding the number of mixture components is key in addressing the heterogeneity across groups. In the context of transplant center evaluation, this is about detecting whether there are subgroups of transplant centers that are underperforming or outperforming the rest. There are two commonly used approaches, the model selection approach (Ishwaran et al., 2001; Woo and Sriram, 2006) and the hypothesis testing approach (Chen et al., 2012). The model selection approach seeks a model to adequately describe the data, while the hypothesis testing approach is used to validate scientific claims. In this paper, we focus on the hypothesis testing approach because it quantifies the confidence of our decisions by providing p-values. Among the many hypotheses that we can test, the most important one is $H_0 : \ C_0 = 1$ vs $H_1 : \ C_0 = 2$, where C_0 is the true number of components. This test

is also referred to as the homogeneity test, since the null hypothesis means all transplant centers are from the same homogeneous population with no anomalies. Chen et al. (2012) provided more examples where different orders of mixture models have different scientific interpretations that require testing.

Even though hypothesis tests are not designed for model selection, they can nevertheless be used for such a purpose in an exploratory study. If $H_0: C_0 = 1$ is rejected, we can sequentially test other hypotheses of the form $H_0: C_0 = C$ vs $H_1: C_0 = C+1, C = 2, 3, \ldots$, in search for the true number of components. This is obviously not a consistent model selection procedure, since we have a fixed chance of failing to reject a hypothesis. On the other hand, one can also argue that many widely used model selection procedures are not consistent, such as the Akaike information criterion. In our simulation studies, we show that the sequential test procedure that we propose can vastly outperform the Bayesian information criterion in model selection.

Due to the loss of strong identifiability for finite Gaussian mixture models, the regular asymptotic theory for likelihood ratio tests (LRT) does not hold. Instead, Chen et al. (2012) and Kasahara and Shimotsu (2015) proposed a locally restricted likelihood ratio test that confines the parameter space in a local alternative model to ensure the existence of an asymptotic distribution for the test statistic. We extend such a test to the proposed latent Gaussian mixture models.

2.3.2 Homogeneity test

We first consider $H_0: C_0 = 1$ vs $H_1: C_0 = 2$. We refer to the model under the null hypothesis as the reduced model and the one under the alternative as the full model. When the null hypothesis is true, γ_i are i.i.d. random variables following Normal(μ_0, σ_0^2). However, this model is not uniquely parameterized in the full model, unless we restrict the values of some parameters. Following Chen et al. (2012), we restrict the parameter space under the full model to $\bar{\Theta}_2(\tau) = \{\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi_1, \pi_2)^{\mathrm{T}};\ \mu_1, \mu_2 \in \mathbb{R}, \sigma_1, \sigma_2 \geq 0, \pi_1 = \tau, \pi_2 = 1 - \tau\}$, for a fixed $\tau \in (0, 0.5]$. By doing so, we do not impose any constraints

on the order between μ_1 and μ_2. In $\bar{\Theta}_2(\tau)$, the null model is uniquely parameterized by $\boldsymbol{\theta}_0(\tau) = \{\boldsymbol{\theta}_{y,0}^{\mathrm{T}}, \boldsymbol{\theta}_{\gamma,0}^{\mathrm{T}}(\tau)\}^{\mathrm{T}}$, where $\boldsymbol{\theta}_{\gamma,0}(\tau) = (\mu_0, \mu_0, \sigma_0^2, \sigma_0^2, \tau, 1-\tau)^{\mathrm{T}}$.

Let $\bar{\Theta}_1$ be the parameter space when $C_0 = 1$ and the reduced model estimator be $\widehat{\boldsymbol{\theta}}_{red} = \arg\max_{\boldsymbol{\theta} \in \bar{\Theta}_1} l_{pen}(\boldsymbol{\theta})$, which is the usual maximum likelihood estimator for a GLMM under the Gaussian random effect assumption. Under the full model, the estimator under a fixed τ is $\widehat{\boldsymbol{\theta}}_{full}(\tau) = \arg\max_{\boldsymbol{\theta} \in \bar{\Theta}_2(\tau)} l_{pen}(\boldsymbol{\theta})$. This estimator can be obtained using the EM algorithm described in Section 2.2.2 without the step for updating π_c's. The following proposition provides the convergence rate of $\widehat{\boldsymbol{\theta}}_{full}(\tau)$ under the null hypothesis.

Proposition 2.2 *Under $H_0 : C_0 = 1$ and Assumptions 1-7 in the supplementary material, for any fixed $\tau \in (0, 0.5]$, $\widehat{\boldsymbol{\beta}}_{full}(\tau) - \boldsymbol{\beta}_0 = O_p(n^{-1/2})$, $\widehat{\mu}_{c,full}(\tau) - \mu_0 = O_p(n^{-1/8})$ and $\widehat{\sigma}^2_{c,full}(\tau) - \sigma_0^2 = O_p(n^{-1/4})$ for $c = 1, 2$ where $\widehat{\boldsymbol{\beta}}_{full}$, $\widehat{\mu}_{c,full}$ and $\widehat{\sigma}^2_{c,full}$ are components in $\widehat{\boldsymbol{\theta}}_{full}(\tau)$ while $\boldsymbol{\beta}_0$, μ_0 and σ_0^2 are the true parameters.*

Remark 1 We use a similar reparameterization similar to that of Kasahara and Shimotsu (2015) in the proof of Proposition 2.2. As shown in the proof, many derivatives of the log likelihood are either exactly zero or have mean zero, and it takes a ninth order Taylor expansion to get a local quadratic approximation for the penalized likelihood. The convergence rate in the proposition means that, for an over-fitted mixture model, the regression coefficient $\boldsymbol{\beta}$ still enjoys the root-n convergence rate, while the parameters of the latent Gaussian mixture model converge much slower. This slow convergence rate also stresses a fundamental difference between our latent Gaussian mixture model and the common parametric models. The $O_p(n^{-1/8})$ convergence rate in $\widehat{\mu}_{c,full}$ is in agreement with the minimax lower bound established in Ho and Nguyen (2016) for finite Gaussian mixture models with one redundant component.

For any subset of numbers $\boldsymbol{\mathcal{T}}$ in $(0, 0.5]$, define the test statistic

$$\widetilde{T}_1 = \max_{\tau \in \mathcal{T}} T_1(\tau) \qquad \text{where } T_1(\tau) = 2[l_n\{\widehat{\boldsymbol{\theta}}_{full}(\tau)\} - l_n(\widehat{\boldsymbol{\theta}}_{red})]. \tag{2.5}$$

Proposition 2.3 *Under $H_0 : C_0 = 1$ and Assumptions 1-7, $\widetilde{T}_1 \to \chi^2(2)$ in distribution as $n \to \infty$.*

Remark 2 Our proof of Proposition 2.3 shows that, under $H_0 : C_0 = 1$, $T_1(\tau) \to \chi^2(2)$ *in distribution for any fixed* τ. *In fact, if there is only one true component, no matter how we choose to split that component, the leading term in the asymptotic expansion of* $T_1(\tau)$ *remains the same. We define* $\widetilde{T}_1$ *as the maximum of* $T_1(\tau)$ *over* $\mathcal{T}$ *to increase the power: if* H_1 *is true, the more values of* τ *we try, the better chance we have to detect an extra component. Proposition 2.3 holds if* $\widetilde{T}_1$ *is the maximum of* $T_1(\tau)$ *over the whole interval* $(0, 0.5]$, *but for practical consideration* $\mathcal{T}$ *is often taken as a finite subset. The condition* $a_n = o_p(n^{1/4})$ *also guarantees that the asymptotic distribution of test statistic is not affected by penalty (2.4) in estimation.*

The detailed test procedure is as follows.

Step 0 Obtain $\widehat{\boldsymbol{\theta}}_{red}$ *and* $l_n(\widehat{\boldsymbol{\theta}}_{red})$.

Step 1 For a fixed τ, *obtain* $\widehat{\boldsymbol{\theta}}_{full}(\tau)$. *To guarantee that a global maximum of the penalized likelihood is reached, try 100 randomly selected initial values for* $\boldsymbol{\theta}(\tau)$.

Step 2 (Optional) Using $\widehat{\boldsymbol{\theta}}_{full}(\tau)$ *obtained in Step 1 as the starting value, perform two more EM iterations without fixing* τ, *and use the resulting estimator to evaluate* $T_1(\tau)$.

Step 3 Repeat Steps 1 and 2 for each $\tau \in \mathcal{T}$ *to obtain* $\widetilde{T}_1$, *where* $\mathcal{T}$ *is set to be* $\{0.1, 0.3, 0.5\}$ *following the recommendation of Chen et al. (2012).*

Step 4 For a size α *test, reject* $H_0 : C_0 = 1$ *if* $\widetilde{T}_1 > \chi^2_\alpha(2)$.

In Step 2, we perform two more EM iterations without fixing τ to increase the power of the test, as recommended by Chen et al. (2012).

2.3.3 Testing for C greater than 1

Next, we consider a test $H_0 : C_0 = C$ vs $H_1 : C_0 = C + 1$ for a $C \geq 2$. We now refer to the model with C components as the reduced model and the one with $C+1$ components as the full model. We first compute the reduced model estimator $\widehat{\boldsymbol{\theta}}_{red} = \arg\max_{\boldsymbol{\theta} \in \bar{\Theta}_C} l_{pen}(\boldsymbol{\theta})$. Assuming H_0 is true, denote the true value of the parameter by $\boldsymbol{\theta}_0$ and order the true mean

parameters by $\mu_{1,0} < \cdots < \mu_{C,0}$. This parameter is not uniquely identified in the full model: if any $\pi_c = 0$ or $(\mu_c, \sigma_c) = (\mu_{c+1}, \sigma_{c+1})$ for some $c \in \{1, \ldots, C\}$, the full model degenerates to the reduced model. In order to make the reduced model identifiable in $\bar{\Theta}_{C+1}$, we will impose constraints that $\pi_c > 0$ for all $c = 1, \ldots, C+1$ and $\pi_c/(\pi_c + \pi_{c+1}) = \tau$ for some c and a fixed $\tau \in (0, 0.5]$ like we did in Section 2.3.2.

To test if a $(C+1)$-component mixture model fits the data better, we will test to see if any one of the C components in the reduced model can be further split into two. Define non-overlapping intervals $D_1, \ldots, D_C$ such that $\mu_{c,0} \in D_c$. For a fixed $\tau \in (0, 0.5]$ and $c \in \{1, \ldots, C\}$, define neighborhoods in the parameter space $\bar{\Theta}_{C+1}$: $\mathcal{N}_{C+1}(c, \tau) = \big\{\boldsymbol{\theta} \in \bar{\Theta}_{C+1} \mid \pi_c/(\pi_c + \pi_{c+1}) = \tau;\ \mu_{c'} \in D_{c'}$ for $c' < c$; $\mu_c, \mu_{c+1} \in D_c$; $\mu_{c'} \in D_{c'-1}$ for $c' > c+1\big\}$. The neighborhood $\mathcal{N}_{C+1}(c, \tau)$ collects the parameters that split the cth component into two daughter components with a split proportion τ, while restricting the other mean parameters from changing too much. The definition of $\mathcal{N}_{C+1}(c, \tau)$ requires knowledge about intervals $\{D_1, \ldots, D_C\}$ that contain the true mean parameters. In practice, we already have a consistent estimator of $\mu_{c,0}$ from fitting the reduced model. Replacing $\{D_c\}_{c=1}^C$ with their consistent estimates does not affect the asymptotic behavior of the test we are about to propose. A practical choice for $\{D_c\}_{c=1}^C$ is provided below in the test procedure. Like in Section 2.3.2, we do not restrict order between μ_c and μ_{c+1} in $\mathcal{N}_{C+1}(c, \tau)$ because τ is restricted to $(0, 0.5]$.

Define the locally restricted full model estimator as

$$\widehat{\boldsymbol{\theta}}_{full}(c, \tau) = \arg \max_{\boldsymbol{\theta} \in \mathcal{N}_{C+1}(c,\tau)} l_{pen}(\boldsymbol{\theta}).$$

To obtain this estimator, we need some minor adjustments to the EM algorithm in Section 2.2.2. First, we update $\pi_c + \pi_{c+1}$ as a single parameter and then assign values for π_c and π_{c+1} proportional to τ. Second, after each M-step, we enforce the restrictions in $\mathcal{N}_{C+1}(c, \tau)$ by forcing any $\mu_{c'}$ stepping out of the boundary back to its predetermined range. A similar scheme was used in Chen et al. (2012). The following convergence rate result echoes Proposition 2.2. It shows that the component that we are trying to split

suffers a slower convergence rate, because it is overfitted in $\mathcal{N}_{C+1}(c,\tau)$ as a mixture of two daughter components, and the rest of the parameters converge in root-n rate.

Proposition 2.4 *Under $H_0 : C_0 = C$ and Assumptions 1-8 in the supplementary material, for any fixed $\tau \in (0, 0.5]$,*

$$\widehat{\mu}_{c,full}(c,\tau) - \mu_{c,0} = O_p(n^{-1/8}), \quad \widehat{\mu}_{c+1,full}(c,\tau) - \mu_{c,0} = O_p(n^{-1/8}),$$
$$\widehat{\sigma}^2_{c,full}(c,\tau) - \sigma^2_{c,0} = O_p(n^{-1/4}), \quad \widehat{\sigma}^2_{c+1,full}(c,\tau) - \sigma^2_{c,0} = O_p(n^{-1/4}),$$

and

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_{y,full}(c,\tau) - \boldsymbol{\theta}_{y0} &= O_p(n^{-1/2}), \\
\widehat{\boldsymbol{\theta}}_{\gamma,c',full}(c,\tau) - \boldsymbol{\theta}_{\gamma,c',0} &= O_p(n^{-1/2}) \text{ for } c' < c, \\
\widehat{\boldsymbol{\theta}}_{\gamma,c',full}(c,\tau) - \boldsymbol{\theta}_{\gamma,c'-1,0} &= O_p(n^{-1/2}) \text{ for } c' > c+1,
\end{aligned}$$

where $\boldsymbol{\theta}_{\gamma,c'} = (\mu_{c'}, \sigma^2_{c'}, \pi_{c'})^{\mathrm{T}}$.

To test if any component in the reduced model can be further divided into two, define the test statistic

$$T_C(\tau) = \max_{c\in\{1,\ldots,C\}} T_C(c,\tau), \quad \text{where } T_C(c,\tau) = 2[l_n\{\widehat{\boldsymbol{\theta}}_{full}(c,\tau)\} - l_n(\widehat{\boldsymbol{\theta}}_{red})]. \tag{2.6}$$

For any finite subset of $(0, 0.5]$ $\mathcal{T}$, define the test statistic

$$\widetilde{T}_C = \max_{\tau\in\mathcal{T}} T_C(\tau). \tag{2.7}$$

In order to understand the asymptotic behavior of $T_C(c,\tau)$, we adopt the reparameterization of Kasahara and Shimotsu (2015) in $\mathcal{N}_{C+1}(c,\tau)$. Define the new parameter vector as $\boldsymbol{\psi}(c,\tau) = \{\boldsymbol{\theta}_y^{\mathrm{T}}, \boldsymbol{\delta}(c)^{\mathrm{T}}, \boldsymbol{\mu}(c)^{\mathrm{T}}, \boldsymbol{\sigma}^2(c)^{\mathrm{T}}, \lambda_\mu, \lambda_\sigma\}^{\mathrm{T}}$ such that

$$\begin{pmatrix} \mu_c \\ \mu_{c+1} \\ \sigma^2_c \\ \sigma^2_{c+1} \end{pmatrix} = \begin{Bmatrix} \nu_\mu + (1-\tau)\lambda_\mu \\ \nu_\mu - \tau\lambda_\mu \\ \nu_\sigma + (1-\tau)(2\lambda_\sigma - \frac{1+\tau}{3}\lambda^2_\mu) \\ \nu_\sigma - \tau(2\lambda_\sigma + \frac{2-\tau}{3}\lambda^2_\mu) \end{Bmatrix}, \tag{2.8}$$

and

$$\begin{aligned}
\boldsymbol{\delta}(c) &= (\pi_1,\ldots,\pi_{c-1}, \quad \pi_c+\pi_{c+1}, \quad \pi_{c+2},\ldots,\pi_C)^{\mathrm{T}},\\
\boldsymbol{\mu}(c) &= (\mu_1,\ldots,\mu_{c-1}, \quad \nu_\mu, \quad \mu_{c+2},\ldots,\mu_C,\ \mu_{C+1})^{\mathrm{T}},\\
\boldsymbol{\sigma}^2(c) &= (\sigma_1^2,\ldots,\sigma_{c-1}^2, \quad \nu_\sigma, \quad \sigma_{c+2}^2,\ldots,\sigma_C^2,\ \sigma_{C+1}^2)^{\mathrm{T}}.
\end{aligned} \tag{2.9}$$

Denote the new parameter space as $\bar{\Theta}_{\psi,C+1}$ and partition $\boldsymbol{\psi}$ into $(\boldsymbol{\eta}^{\mathrm{T}},\boldsymbol{\lambda}^{\mathrm{T}})^{\mathrm{T}}$ where

$$\boldsymbol{\eta} = \{\boldsymbol{\theta}_y^{\mathrm{T}},\boldsymbol{\delta}(c)^{\mathrm{T}},\boldsymbol{\mu}(c)^{\mathrm{T}},\boldsymbol{\sigma}^2(c)^{\mathrm{T}}\}^{\mathrm{T}}, \quad \boldsymbol{\lambda}=(\lambda_\mu,\lambda_\sigma)^{\mathrm{T}}.$$

The reduced model is uniquely parameterized by $\boldsymbol{\theta}^* \in \mathcal{N}_{C+1}(c,\tau)$, and it is reparameterized as $\boldsymbol{\psi}^* = \{(\boldsymbol{\eta}^*)^{\mathrm{T}},0,0\}^{\mathrm{T}}$, or more specifically $\boldsymbol{\theta}_y = \boldsymbol{\theta}_{y,0}$, $\boldsymbol{\lambda}^* = 0$ and $\boldsymbol{\delta}^*(c) = (\pi_{1,0},\ldots,\pi_{C-1,0})^{\mathrm{T}}$, $\boldsymbol{\mu}^*(c) = (\mu_{1,0},\ldots,\mu_{C,0})^{\mathrm{T}}$, $\boldsymbol{\sigma}^{2*}(c) = (\sigma_{1,0}^2,\ldots,\sigma_{C,0}^2)^{\mathrm{T}}$. The reparameterization in (2.8) is beneficial because, to test if the cth component can be further split, we can equivalently test if $\boldsymbol{\lambda}=0$.

Define

$$\boldsymbol{s}_i^{(c)} = \left\{\boldsymbol{s}_{\boldsymbol{\eta},i}^{\mathrm{T}},(\boldsymbol{s}_{\boldsymbol{\lambda},i}^{(c)})^{\mathrm{T}}\right\}^{\mathrm{T}}, \tag{2.10}$$

where

$$\begin{aligned}
\boldsymbol{s}_{\boldsymbol{\eta},i} &= (\boldsymbol{s}_{\boldsymbol{\theta}_y,i}^{\mathrm{T}},\boldsymbol{s}_{\boldsymbol{\delta},i}^{\mathrm{T}},\boldsymbol{s}_{\mu,i}^{\mathrm{T}},\boldsymbol{s}_{\sigma,i}^{\mathrm{T}})^{\mathrm{T}},\\
\boldsymbol{s}_{\lambda,i}^{(c)} &= \left(\int \zeta_i\pi_c f_{c,i}^* H_{ci}^{3*}/\int \zeta_i g^*, \int \zeta_i\pi_c f_{c,i}^* H_{ci}^{4*}/\int \zeta_i g^*\right)^{\mathrm{T}},\\
\boldsymbol{s}_{\boldsymbol{\theta}_y,i} &= \int (\partial\zeta_i/\partial\boldsymbol{\theta}_y) g^*/\int \zeta_i g^*,\\
\boldsymbol{s}_{\boldsymbol{\delta},i} &= \{\int \zeta_i(f_{1,i}^*-f_{C,i}^*)/\int \zeta_i g_i^*,\ldots,\int \zeta_i(f_{C-1,i}^*-f_{C,i}^*)/\int \zeta_i g_i^*\}^{\mathrm{T}},\\
\boldsymbol{s}_{\mu,i} &= (\int \zeta_i\pi_1 f_{1,i}^* H_{1i}^{1*}/\int \zeta_i g^*,\ldots,\int \zeta_i\pi_C f_{C,i}^* H_{Ci}^{1*}/\int \zeta_i g^*)^{\mathrm{T}}
\end{aligned}$$

and

$$\boldsymbol{s}_{\sigma,i} = (\int \zeta_i\pi_1 f_{1,i}^* H_{1i}^{2*}/\int \zeta_i g^*,\ldots,\int \zeta_i\pi_C f_{C,i}^* H_{Ci}^{2*}/\int \zeta_i g^*)^{\mathrm{T}}.$$

Here, we use the short hand notation $\zeta_i = \prod_{k=1}^{N_i} f(y_{ik} \mid \boldsymbol{x}_{ik},\gamma_i;\boldsymbol{\theta}_y)$, $f_{c,i}^* = f_c(\gamma_i \mid \mu_{c,0},\sigma_{c,0})$, $g_i^* = g(\gamma_i \mid \boldsymbol{\theta}_\gamma^*)$ and $H_{ci}^{k*} = H^k\{(\gamma_i-\mu_{c,0})/\sigma_{c,0}\}/(k!\sigma_{c,0}^k)$, where $H^k(\cdot)$ is the kth Hermite Polynomial.

Proposition 2.5 *Under $H_0 : C_0 = C$ and Assumptions 1-8 in the Appendix,*

$$\widetilde{T}_C \to \max\left\{(\boldsymbol{S}^{(c)}_{\lambda|\eta,n})^T(\boldsymbol{\mathcal{I}}^{(c)}_{\lambda|\eta})^{-1}\boldsymbol{S}^{(c)}_{\lambda|\eta,n}, c = 1,\ldots,C\right\} \text{ in distribution,}$$

where $\boldsymbol{S}^{(c)}_{\lambda|\eta,n} = \boldsymbol{S}^{(c)}_{\lambda,n} - \boldsymbol{\mathcal{I}}^{(c)}_{\lambda\eta}\boldsymbol{\mathcal{I}}^{-1}_{\eta}\boldsymbol{S}_{\eta,n}$, $\boldsymbol{\mathcal{I}}^{(c)}_{\lambda|\eta} = \boldsymbol{\mathcal{I}}^{(c)}_{\lambda} - \boldsymbol{\mathcal{I}}^{(c)}_{\lambda\eta}\boldsymbol{\mathcal{I}}^{-1}_{\eta}(\boldsymbol{\mathcal{I}}^{(c)}_{\lambda\eta})^T$, $\boldsymbol{S}_{\eta,n} = n^{-1/2}\sum_{i=1}^n \boldsymbol{s}_{\eta,i}$, $\boldsymbol{S}^{(c)}_{\lambda,n} = n^{-1/2}\sum_{i=1}^n \boldsymbol{s}^{(c)}_{\lambda,i}$, $\boldsymbol{\mathcal{I}}^{(c)}_{\lambda\eta} = E\{\boldsymbol{s}^{(c)}_{\boldsymbol{\lambda},i}\boldsymbol{s}^{\mathrm{T}}_{\boldsymbol{\eta},i}\}$, $\boldsymbol{\mathcal{I}}_{\eta} = E(\boldsymbol{s}_{\eta,n}\boldsymbol{s}^{\mathrm{T}}_{\eta,n})$, $\boldsymbol{\mathcal{I}}^{(c)}_{\lambda} = E\{\boldsymbol{s}^{(c)}_{\boldsymbol{\lambda},i}(\boldsymbol{s}^{(c)}_{\boldsymbol{\lambda},i})^{\mathrm{T}}\}$.

One can show $(\boldsymbol{S}^{(c)}_{\lambda|\eta,n})^T(\boldsymbol{\mathcal{I}}^{(c)}_{\lambda|\eta})^{-1}\boldsymbol{S}^{(c)}_{\lambda|\eta,n} \to \chi^2(2)$ in distribution for each c, but the score vectors $\boldsymbol{S}^{(c)}_{\lambda|\eta,n}$ are correlated across different c's and hence the distribution of $\widetilde{T}_C$ in Proposition 2.5 is that of the maximum of a few correlated $\chi^2(2)$ random variables. In the supplementary material, we describe a simulation method to evaluate this asymptotic distribution. This procedure only requires estimating the covariance matrix of $\{\boldsymbol{S}^{(c)}_{\lambda|\eta,n}, c = 1,\ldots,C\}$ and simulating Gaussian random variables. It is extremely fast and fundamentally different from bootstrap, which requires fitting the model a large number of times to the bootstrap samples.

For any $C \geq 2$, our test procedure for $H_0 : C_0 = C$ is as follows.

Step 0 Obtain $\widehat{\boldsymbol{\theta}}_{red}$ and evaluate $l_n(\widehat{\boldsymbol{\theta}}_{red})$. Define subintervals

$$\begin{aligned} D_1 &= [\widehat{\gamma}_{min},\ \widehat{\mu}_{1,red}/2 + \widehat{\mu}_{2,red}/2], \\ D_2 &= (\widehat{\mu}_{1,red}/2 + \widehat{\mu}_{2,red}/2,\ \widehat{\mu}_{3,red}/2 + \widehat{\mu}_{2,red}/2], \\ &\vdots\ , \\ D_C &= (\widehat{\mu}_{C-1,red}/2 + \widehat{\mu}_{C,red}/2,\ \widehat{\gamma}_{max}], \end{aligned}$$

where $\widehat{\gamma}_{min}$ and $\widehat{\gamma}_{max}$ are the minimum and maximum of the predicted γ's under the reduced model.

Step 1 Obtain $\widehat{\boldsymbol{\theta}}_{full}(c,\tau)$ by maximizing the penalized likelihood of the full model in the restricted parameter neighborhood $\mathcal{N}_{C+1}(c,\tau)$ using the subintervals $\{D_k\}_{k=1}^C$ defined in Step 0. The penalty on σ_k^2 is $p_n(\sigma_k^2, \widehat{\sigma}^2_{c',red})$ if μ_k is restricted in $D_{c'}$ for $k = 1,\ldots,C+1$ and a_n is chosen according equation (23) in Kasahara and Shimotsu (2015). If a μ_k steps outside of its range $D_{c'}$ specified by $\mathcal{N}_{C+1}(c,\tau)$ during the EM iterations, we simply set it back to the nearest boundary of $D_{c'}$. To ensure that the maximum of l_{pen} is reached, we repeat the EM algorithm 100 times using randomly selected initial values within $\mathcal{N}_{C+1}(c,\tau)$.

Step 2 Using $\widehat{\boldsymbol{\theta}}_{full}(c,\tau)$ *as the starting value, do two more EM iterations without fixing* τ*. Use the resulting estimator to evaluate* $T_C(c,\tau)$ *in (2.6).*

Step 3 Repeat Steps 1 and 2 for each $c \in \{1,\ldots,C\}$ *and* $\tau \in \mathcal{T} = \{0.1, 0.3, 0.5\}$*, and evaluate* $\widetilde{T}_C$ *in (2.7).*

Step 4 Evaluate the null distribution in Proposition 2.5 using the procedure described in the supplementary material and compare $\widetilde{T}_C$ *with the null distribution to get the p value.*

2.4 Use False Discovery Rate Control to Classify Groups

A practical utility of model (2.1) is to classify groups based on γ_i. To ease understanding, we frame the ensuing development in the context of the aforementioned transplant center evaluation. That is, different components in the mixture density $g(\gamma)$ represent different clusters in health care quality delivered by transplant centers, and we want to classify the transplant centers into these clusters. However, these clusters are not considered to be equal: usually a subset of clusters, denoted as $\mathcal{C}_0$, represent the norm of care quality, consisting of centers with average performances; those out of $\mathcal{C}_0$ are centers either underperforming or outperforming the industrial standard. Following Efron's "empirical null" idea (Efron, 2004), $\mathcal{C}_0 \subset \{1,\ldots,C\}$ can be identified as one or more components in the fitted mixture model, usually those in the middle of $g(\gamma)$ with high weights π_c's.

With $\mathcal{C}_0$ representing the distribution of normal care quality, one should classify an individual center into clusters outside of $\mathcal{C}_0$ with extreme care, since it declares that center as an anomaly, and the false discovery rate needs to be controlled. As pointed out in Sun et al. (2015), classification problems with unequal losses in different classes are naturally connected with multiple hypothesis tests. In our context, this classification problem is equivalent to performing a test for each center on whether the center is in the empirical null $\mathcal{C}_0$. In other words, we test a sequence of hypotheses $H_{i0} : \sum_{c\in\mathcal{C}_0} L_{ic} = 1$, $i = 1,\ldots,n$. Since $\mathcal{C}_0$ represents the average quality of care, center i is considered "interesting" (either outperforming or underperforming) if H_{i0} is rejected.

For a given subset of components $\mathcal{C}_0$, identify the "empirical null" distribution of γ as $g_0(\gamma \mid \boldsymbol{\theta}_\gamma) = \sum_{c\in\mathcal{C}_0} \pi_c f_c(\gamma \mid \mu_c, \sigma_c)/\sum_{c\in\mathcal{C}_0} \pi_c$. Since γ_i is not directly observed, our decision rule for H_{i0} is based on the observed data $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$, denoted as $\delta_i = \delta(\boldsymbol{X}_i, \boldsymbol{Y}_i; \boldsymbol{\theta})$, where $\delta_i = 1$ means center i is "interesting" and $\delta_i = 0$ otherwise. The false discovery rate is defined as

$$FDR = E\left\{ \frac{\sum_i^n I(\delta_i = 1, \sum_{c\in\mathcal{C}_0} L_{ic} = 1)}{\sum_i^n I(\delta_i = 1)} \mid \sum_i^n I(\delta_i = 1) > 0 \right\} pr\left\{ \sum_i^n I(\delta_i = 1) > 0 \right\}$$

When γ_i's are observed, Sun and Cai (2007) show that the oracle decision rule is based on the local FDR, $T_{\mathrm{OR}}(\gamma_i) = P(\sum_{c\in\mathcal{C}_0} L_{ic} = 1 \mid \gamma_i) = \sum_{c\in\mathcal{C}_0} \pi_c f_c(\gamma_i)/g(\gamma_i)$. In our case, γ_i is not observed, and the local FDR is defined as the posterior probability given the observed data

$$lFDR_i = pr(\textstyle\sum_{c\in\mathcal{C}_0} L_{ic} = 1 \mid \boldsymbol{X}_i, \boldsymbol{Y}_i) = \dfrac{\sum_{c\in\mathcal{C}_0} \pi_c \int f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma; \boldsymbol{\beta}) f_c(\gamma \mid \mu_c, \sigma_c) d\gamma}{\int f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma; \boldsymbol{\beta}) g(\gamma \mid \boldsymbol{\theta}_\gamma) d\gamma}. \quad (2.11)$$

It is easy to show $lFDR_i = E\{T_{\mathrm{OR}}(\gamma_i) \mid \boldsymbol{X}_i, \boldsymbol{Y}_i\}$. Following Sun et al. (2015), the multiple hypothesis testing problem is related to a classification problem with the loss function

$$\mathscr{L}(\boldsymbol{L}, \boldsymbol{\delta}) = \lambda \textstyle\sum_i \delta_i (\sum_{c\in\mathcal{C}_0} L_{ic}) + \sum_i (1 - \delta_i)(1 - \sum_{c\in\mathcal{C}_0} L_{ic}),$$

where λ is a penalty for false positives. Let $\mathscr{R} = E\{\mathscr{L}(\boldsymbol{L}, \boldsymbol{\delta})\}$ be the risk of the classification problem. By Theorem 1 of Sun et al. (2015), the optimal decision rule that minimizes this risk is $\delta_i = I(lFDR_i < t)$ for some threshold t.

Let $lFDR_{(1)} \le \cdots \le lFDR_{(n)}$ be the ranked lFDR values. For any $\alpha > 0$, let

$$k = \max_i \{ \frac{1}{i} \sum_{j=1}^{i} lFDR_{(j)} \le \alpha \}$$

and our FDR control procedure is to reject all H_{i0} with the rank of $lFDR_i$ less or equal to k.

Proposition 2.6 *Under the model in (2.1), the above procedure controls FDR at level α.*

A sketch proof of Proposition 2.6 is provided in Section S.8 of the supplementary material. In practice, $lFDR$ is estimated by substituting $\boldsymbol{\theta}$ with its estimator and the integrals in (2.11) are evaluated using Gaussian quadrature as described in the supplementary material.

2.5 Simulation Studies

2.5.1 Simulation 1: Estimation and random effect prediction

We simulate data for $n = 282$ transplant centers, which is the number of kidney transplant centers in the Organ Procurement and Transplantation Network in the year 2008. The number of patients per center has a highly skewed distribution in the real data. To mimic such a distribution, we generate N_i as the floor of the sum of $Poisson(5)$ and $Exponential(45)$. The response Y_{ik} is a binary variable generated using (2.1) with $P(Y_{ik} = 1) = \{1 + \exp(-\xi_{ik})\}^{-1}$, where $\xi_{ik} = \boldsymbol{X}_{ik}^{\mathrm{T}}\boldsymbol{\beta} + \gamma_i$, $\boldsymbol{X}$ is generated from a bivariate standard normal distribution and $\boldsymbol{\beta} = (1, 1)^{\mathrm{T}}$. We generate γ_i's from the following Gaussian mixture models

$$\text{Model 1: } 0.5\ N(-3.26, 1.2^2) + 0.5\ N(0.74, 0.8^2),$$

$$\text{Model 2: } 0.3\ N(-5.26, 1.2^2) + 0.4\ N(-0.26, 0.8^2) + 0.3\ N(2.74, 0.9^2).$$

The parameters in these models are selected such that the marginal probability of $\{Y_{ik} = 1\}$ for each model is roughly the same as for the real data. We repeat the simulation 200 times under each model and apply the estimation procedure in Section 2.2.2 to each simulated data set. Estimation results for Model 1 and Model 2 in Simulation 1, under correctly specified number of components, are summarized in Table 2.1 and 2.2 respectively. The mixture components in the estimated model are ranked according to the value $\widehat{\mu}_c$ to avoid the cluster label switching problem.

We can see that the estimation results are quite reasonable: all biases are virtually zero; the standard errors for component means (μ_c) and component standard deviations (σ_c) are slightly inflated compared with Table 2.1, which is understandable since we are fitting a more complicated mixture model; the standard errors for $\boldsymbol{\beta}$ are not affected by the increased complicity of the latent mixture model.

To illustrate the consequence for mis-specifying the random effect distribution, we also fit a common GLMM to the simulated data under the assumption that γ_i's are i.i.d. Gaussian. In Table 2.3, we report the mean square prediction error for the random effect averaged over the 200 simulation runs and the Monte Carlo standard deviation of the prediction

Table 2.1 Summary for parameter estimation under Simulation Model 1 based on 200 replications.

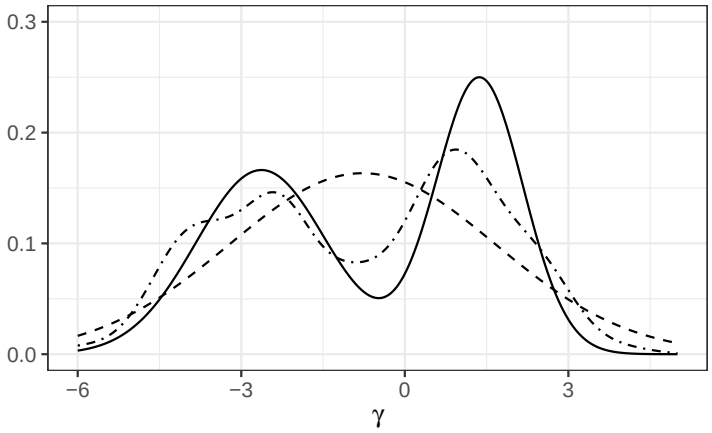
	Truth	Mean	Bias	Std
π_1	0.5000	0.4971	-0.0029	0.0280
π_2	0.5000	0.5029	0.0029	0.0280
μ_1	-3.2598	-3.2586	0.0012	0.1262
μ_2	0.7402	0.7401	-0.0001	0.0752
σ_1	1.2000	1.1954	-0.0046	0.1340
σ_2	0.8000	0.7960	-0.0040	0.0630
β_1	1.0000	1.0017	0.0017	0.0213
β_2	1.0000	1.0006	0.0006	0.0225

Table 2.2 Summary for parameter estimation under Simulation Model 2 based on 200 replications.

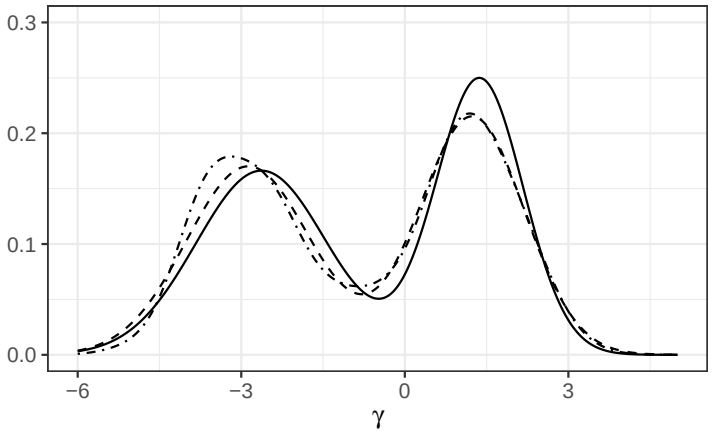
	Truth	Mean	Bias	Std
π_1	0.3000	0.3016	0.0016	0.0244
π_2	0.4000	0.3904	-0.0096	0.0588
π_3	0.3000	0.3080	0.0080	0.0596
μ_1	-5.2598	-5.2800	-0.0202	0.2175
μ_2	-0.2598	-0.2652	-0.0054	0.3472
μ_3	2.7402	2.6894	-0.0508	0.3433
σ_1	1.2000	1.1821	-0.0179	0.2664
σ_2	0.8000	0.8036	0.0036	0.1948
σ_3	0.9000	0.9286	0.0286	0.2516
β_1	1.0000	1.0010	0.0010	0.0225
β_2	1.0000	1.0038	0.0038	0.0226

error. As we can see, when the random effect distribution is mis-specified as Gaussian, the fitted model yields a much larger prediction error. Figure 2.1 illustrates the effect of model misspecification on random effect prediction. The data are generated in a typical simulation run under simulation Model 1. The upper panel shows the prediction results of a common generalized linear mixed model under Gaussian random effect assumption, and the lower panel shows the results of the proposed model. In both panels, we compare the true density of γ with the estimated density using the fitted model and the kernel density of the predicted γ using the fitted model. As we can see from the upper panel, prediction under the mis-specified Gaussian random effect assumption suffers from a shrinkage effect that the values of $\widehat{\gamma}$ are pushed towards the center of the distribution so that the posterior

distribution resembles the shape of a Gaussian distribution. The lower panel shows that prediction under our proposed model does not suffer from such a shrinkage effect. Our model recovers the shape of the latent variable distribution and produces better predictions.



(a)



(b)

Figure 2.1 Simulation Model 1: impact of random effect assumption. Panel (a) shows results from a common generalized linear mixed model with a mis-specified Gaussian random effect assumption; Panel (b) shows results of the proposed latent Gaussian mixture model with a correctly specified number of components. In both panels, the solid curve is the true density for γ, the dashed curve is the estimated density of γ using the fitted model, and the dot-dash curve is the kernel density of the predicted random effects.

2.5.2 Simulation 2: Hypothesis tests

Next, we investigate the validity and power for the proposed tests in Section 2.3. We generate simulated data under similar settings as in Simulation 1, while γ_i's are generated

Table 2.3 Mean squared prediction error for the random effect under Simulation Models 1 and 2. Gaussian: GLMM with Gaussian random effects; Gaussian Mixture: the proposed model; Mean: Mean Squared Prediction Error averaged over 200 replicates; Std: standard deviation of the prediction error.

Simulation Model	Fitted Model	Mean	Std
Model 1	Gaussian	0.4167	0.0392
	Gaussian Mixture	0.3589	0.0361
Model 2	Gaussian	0.6988	0.0697
	Gaussian Mixture	0.5405	0.0581

from three models: Model 1, Model 2 and

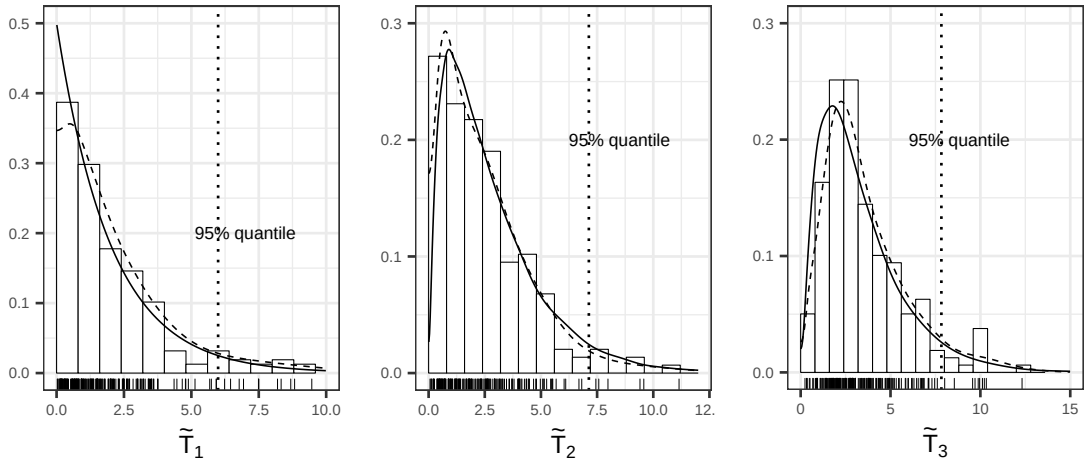
$$\text{Model 0: } N(-1.26, 0.5^2).$$

The three models represent latent Gaussian mixture models with orders 1 to 3. We generate 200 simulated data sets under each of the three models, and compute $\widetilde{T}_1$ in data under Model 0, $\widetilde{T}_2$ under Model 1 and $\widetilde{T}_3$ under Model 2. The empirical distributions of the three quantities represent the null distribution for the test statistics under the null hypotheses $C_0 = 1, 2$ and 3 respectively. These empirical distributions are provided in Fig. 2.2 and compared with the asymptotic distributions provided in Section 2.3. In each panel of Fig. 2.2, the dash curve is the kernel density based on 200 replicates of the test statistic and the solid curve is the asymptotic distribution. The asymptotic distributions for $\widetilde{T}_2$ and $\widetilde{T}_3$ are based on 10,000 simulations using the procedure described in the supplementary material. As we can see, the empirical distributions of the test statistics are remarkably close to the asymptotic distribution, which also shows the validity of the proposed tests. We use $\widetilde{T}_1$ - $\widetilde{T}_3$ to test the three null hypotheses, and the empirical sizes of these tests are 0.06, 0.03 and 0.05 respectively, which are close to the nominal level 0.05.

Next, we illustrate the power of the tests. The response Y is generated the same way as in Section 2.5.1, while γ is generated from the following two models:

$$\text{Model 3: } 0.6\ N(-2.26, 1.2^2) + 0.4\ N(-0.46, 0.8^2),$$

$$\text{Model 4: } 0.3\ N(-3.26, 1.2^2) + 0.4\ N(-0.26, 0.8^2) + 0.3\ N(2.34, 0.9^2).$$



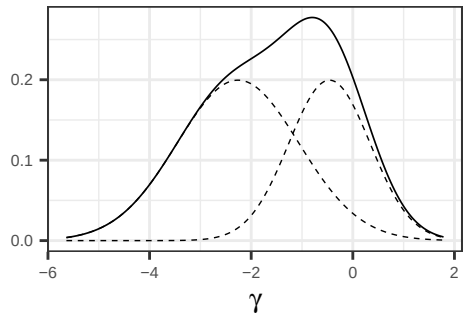
(a) Null distribution of $\widetilde{T}_1$ (b) Null distribution of $\widetilde{T}_2$ (c) Null distribution of $\widetilde{T}_3$

Figure 2.2 Empirical (dash) and asymptotic (solid) distributions of $\widetilde{T}_1$, $\widetilde{T}_2$ and $\widetilde{T}_3$ under the null hypotheses. The vertical dotted line marks the 95% quantile of the asymptotic distribution.

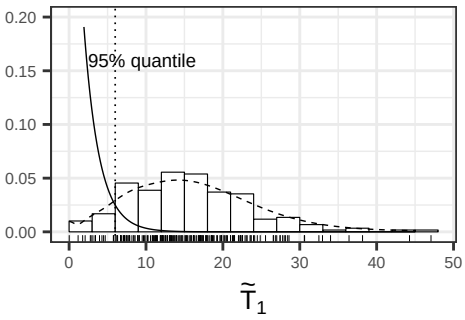
Compared with the Models 1 and 2 considered in Section 2.5.1, the individual components in Models 3 and 4 are less separated, making it harder to detect the real order of these models, especially when γ is an unobserved latent variable.

To examine the power of the proposed locally restricted likelihood ratio tests in Section 2.3, we test $H_0 : C_0 = 1$ when the data are generated from Model 3, and test $H_0 : C_0 = 2$ when the data are generated from Model 4. In Fig. 2.3, we present the empirical distributions of the test statistics based on 200 simulation runs. When performing 5% tests, the empirical powers of the proposed tests are 91% under Model 3 and 95.5% under Model 4. We have also examined the power of the homogeneity test when γ_i's are simulated from Model 1 and the power of the test on $H_0 : C_0 = 2$ when γ_i's are generated from Model 2. The power under both of these cases is virtually equal to 1.

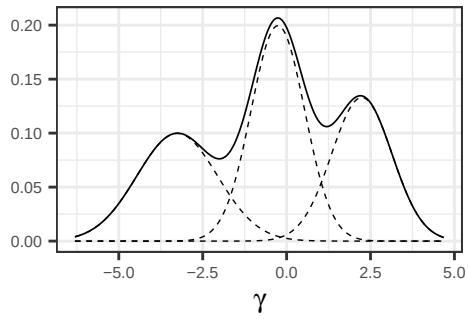
Since a sequential test can be used for model selection purposes, it is of interest to compare the test based procedure with other model selection procedures such as the Bayesian information criterion, which is the negative log likelihood for the observed data plus a penalty on $\log(n)$ times the number of free parameters in the model. For Model 3, the Bayesian information criterion picks the correct model with 2 components in 39% out of



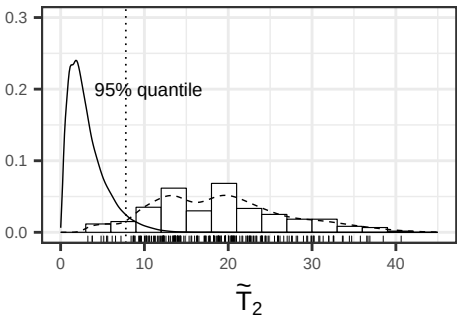
(a) Mixture Model 3



(b) Empirical distribution of $\widetilde{T}_1$



(c) Mixture Model 4



(d) Empirical distribution of $\widetilde{T}_2$

Figure 2.3 Power of the locally restricted likelihood ratio tests. Panels (a) and (c) illustrate the true density (solid) of γ under Model 3 and 4 respectively. The dashed lines represent the individual components. Panels (b) and (d) illustrate the empirical distributions (dash) of $\widetilde{T}_1$ and $\widetilde{T}_2$ comparing to the corresponding null distributions (solid). The vertical dotted line marks the 95% quantile of the null distribution.

the 200 simulations and chooses a 1-component model for the remaining 61% of the repetitions. This means if we use the Bayesian information criterion as the decision rule to test $H_0 : C_0 = 1$ under Model 3, it only has 39% power, which is much lower than the test we developed. For Model 4, the Bayesian information criterion chooses a correct 3-component model in 50.5% of the 200 simulations and chooses 1 or 2 components in the other 49.5% of runs. On the other hand, the sequential test procedure with $\alpha = 0.05$ chooses the correct number of components 88.5% of the time for Model 3, and 86% of the time for Model 4.

2.6 Data Analysis

2.6.1 Background

Renal failure is one of the most common and severe diseases in the United States. In 2013, a total of 117,162 new cases were reported (www.USRDS.org). Kidney transplantation, a primary therapy for end stage renal disease, is a complicated procedure typically involving transplant surgeons and physicians, coordinators, social workers, financial counselors, nutritionists, psychologists and referring physicians. The quality of care delivered by a transplant center is often assessed by patient survival, such as the 5 year post-transplant survival rate.

To provide a fair assessment of each transplant center, both patient level risk factors and an effect representing the quality of care of the transplant center are often included in the risk adjustment model. Many statisticians and health policy researchers model the transplant center effects as random effects that follow a Gaussian distribution (Krumholz et al., 2006a,b; Li et al., 2009). This approach ignores the heterogeneity among the transplant centers, and the assumption of a common Gaussian distribution induces a shrinkage effect that makes the predicted random effects similar in value. He et al. (2013) argue that borrowing information from other transplant centers is not fair when the goal of the study is to evaluate the centers and advocate modeling the transplant center effects as fixed effects. However, in such a fixed effects model, the number of parameters is large, making statistical inference numerically unstable, especially when the center size varies substantially. A comprehensive critique of these two approaches can be found in a report prepared by the Committee of Presidents of Statistical Societies (COPSS) through a contract with Centers for Medicare and Medicaid Services (Ash et al., 2012).

Our proposed latent Gaussian mixture model bridges the gap between the existing approaches and has two advantages. First, the model allows the presence of heterogeneities (e.g. the existence of clusters or subpopulations) among the transplant centers, making it a natural framework to identify centers with anomaly performance. Second, the mixture model can be considered as a compromise between the random effects model and the fixed

effects model: it reduces to the random effects model when there is only one component in the mixture distribution and it becomes the fixed effects model if each transplant center forms a cluster of its own.

Our motivating data are obtained from the Organ Procurement and Transplantation Network, administered by the U.S. Department of Health and Human Services. The data system includes data on all donors, wait-listed candidates, and transplant recipients in the U.S. Included in the analysis are adult renal failure patients ($\geq$ 18 years of age) who underwent deceased donor kidney transplantation between January 1987 and December 2008. This cohort includes $N = 269,386$ patients receiving kidney transplants from a total of $n = 296$ centers. The number of transplants performed by a center, N_i, has a highly skewed distribution. Most centers performed a few hundred cases of kidney transplantation, but there are centers that took over 5000 cases. The patient level response is the 5-year survival status (1=death and -1=survival) and there is no censoring due to routine and rigorous tracking of the patients. The overall 5-year failure rate is 27.59%.

An important patient level covariate that is directly related to the success of kidney transplants is x_1 = cold ischemic time, which is the time that the donor kidney was kept in a refrigerator before being received by the patient. Other patient level covariates include x_2 = age at transplantation and x_3 = sex of the patient (1 = male, 0 = female), while x_4–x_6 are indicators for BMI in the intervals (22, 25], (25, 30] and 30+ respectively. Since the data were collected over a time span of two decades, it is possible that the technology used in transplant surgeries has improved over time, which also affects the patient level outcome. Therefore, in addition to the other covariates described above, we also include time effects into the model. Using cases before 1990 as the baseline, covariates x_7–x_{10} are indicators for cases performed in 1990–1994, 1995–1999, 2000–2003 and 2004–2008 respectively.

2.6.2 Model fitting

We fit the proposed model to the data, using a random effect following a Gaussian mixture distribution to represent the care quality of a center. Using the proposed test procedure to decide the order of the latent Gaussian mixture model, the p-value is 0.0016

for $H_0 : C_0 = 1$ vs. $H_1 : C_0 = 2$; and 0.4076 for $H_0 : C_0 = 2$ vs. $H_1 : C_0 = 3$. We conclude that the care quality among the kidney transplant centers is not homogeneous and and the distribution of the random effect is adequately described by a two-component Gaussian mixture. The estimated fixed effects under our final model are summarized in Table 2.4, where the standard errors are obtained using the asymptotic expansion (S.13) in the supplementary material. As we can see, all covariates considered are significant. Since we code $Y = 1$ as death, the results in Table 2.4 imply that having longer donor kidney delivery times, being older, being male, and having higher BMI all lead to increased risk of patient death. The coefficients for x_7–x_{10} are negative and decreasing, confirming that the overall death rate is decreasing over time.

Table 2.4 U.S. Organ Procurement and Transplantation Network data analysis: estimated fixed effect coefficients, standard errors, z-values and p-values. .

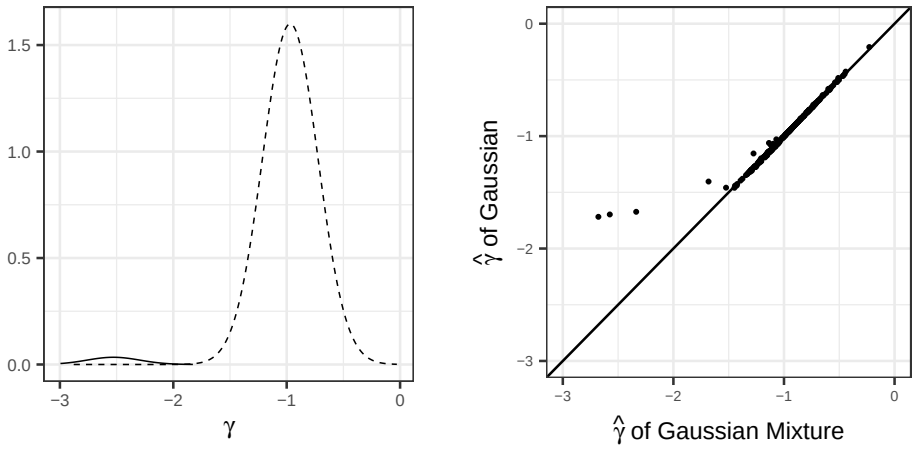
	Estimate	Std. Error	z-value	p-value
x_1	0.019503	0.0003057	63.7996	<1e-99
x_2	0.007112	0.0002118	33.5790	<1e-99
x_3	0.030928	0.0094643	3.2679	0.0011
x_4	0.077860	0.0154823	5.0290	<1e-6
x_5	0.120536	0.0129443	9.3120	<1e-19
x_6	0.225015	0.0147424	15.2632	<1e-51
x_7	-0.270078	0.0145225	-18.5972	<1e-76
x_8	-0.526297	0.0126497	-41.6055	<1e-99
x_9	-0.632073	0.0137602	-45.9348	<1e-99
x_{10}	-0.800276	0.0128928	-62.0714	<1e-99

The estimated Gaussian mixture model for the random effect γ is

$$0.98\ N(-0.969, 0.244^2) + 0.02\ N(-2.528, 0.234^2).$$

The mixture density $g(\gamma)$, as well as its individual components, are illustrated in Fig. 2.4 (a). The majority of the centers have rather similar care quality, but there is also a small cluster of transplant centers that have lower death rates after taking into account all the patient level covariates and these are the centers that are outperforming the others. In Fig. 2.4 (b), we also compare the predicted random effects under the standard GLMM with those under our latent Gaussian mixture model. While the predicted γ is almost the same under both models for the majority of the centers, the care quality effects for the a few

centers in the left tail are severely shrunken towards the mean if we assume the random effects follow a homogeneous Gaussian distribution.



(a) (b)

Figure 2.4 (a) Estimated latent Gaussian mixture model for the kidney transplant data. The solid line and dashed line represent two components. (b) Comparison of the predicted random effects under Gaussian and Gaussian mixture model assumptions.

Since the second component is small, we also run additional simulations to confirm that our methodology really works under such situations. To mimic the real data, we simulate binary Y_{ik} from a logistic GLMM using the covariates from the real data, set $\boldsymbol{\beta}$ as the estimated values in Table 2.4 and generate $\boldsymbol{\gamma}$ from the following mixture model:

$$(1-\pi_2)N(-0.969, 0.244^2) + \pi_2 N(-2.528, 0.234^2).$$

We set π_2 to be 0.005, 0.01, 0.02 or 0.05, and simulate 200 data sets under each setting. The empirical powers for testing $H_0 : C_0 = 1$ are 47%, 78.5%, 97.5% and 100% respectively. These results show that our method can detect a small component under the sample size of the real data and our discovery is likely to be true.

2.6.3 Performance evaluation

Based on the fitted model for γ in Fig. 2.4 (a), the majority of the centers provide similar care for their patients. However, the smaller mixture component consists of transplant centers with lower adjusted mortality rates, and these centers outperform the rest. We let

the empirical null distribution be the bigger component of the fitted mixture model. Using the evaluation procedure described in Section 2.4, we find three transplant centers that outperform the rest. In Table 2.5, we list the IDs of the three outperforming centers, as well as their $lFDR$, $\widehat{\gamma}$, number of cases treated, and average 5-year survival rate.

Table 2.5 The outperforming centers detected using local false discovery rate in the kidney transplant data.

Center ID	lFDR	$\hat{\gamma}$	Sample Size	Survival Rate
#287	0.0013	-2.6784	114	0.973
#10	0.0061	-2.5753	125	0.944
#28	0.0736	-2.3364	120	0.841

2.7 Summary

We propose a GLMM model with latent Gaussian mixture random effects that provides a natural framework to model the inhomogeneity among transplant centers and to rank their care quality. We demonstrate that the predicted random effects can be severely shrunken toward the mean if the distribution of the random effect is mis-specified as Gaussian. This shrinkage effect is quite prominent for the centers in the tails of the population. The latent Gaussian mixture model is not strongly identifiable and suffers from a slow convergence rate when the number of mixture components is larger than the truth. We develop test procedures to decide the number of mixture components. Even though the proposed tests are designed mainly for testing scientific claims and providing uncertainty assessments, they can also be used for model selection and our simulation results in Section 2.5.2 suggest that the sequential test procedure outperforms a naive Bayesian information criterion. We leave development of a consistent model selection procedure for the latent Gaussian mixture model for future work. The proposed test procedures are computationally intensive, especially when analyzing large medical data sets like the OPTN data, since we have to try hundreds of initial values to find the biggest likelihood ratio. These computations are best handled using parallel computing. We have developed a software package `LatentGaussianMixtureModel` written in Julia (http://julialang.org/), which is a high-level, high-performance dynamic

programming language. Our package is based on open source math libraries and supports parallel computing. We will make the package available on the corresponding author's website. Even though comparing transplant centers using the five-year survival rates of the patients has been the standard in the health policy literature, we acknowledge the fact that survival time is a more informative response variable. We intend to explore extending the latent Gaussian mixture model to survival outcomes in future research.

Supplementary Material The online supplementary material contains the model fitting algorithm, additional simulation results and theoretical proofs.

CHAPTER 3. SUBGROUP ANALYSIS AND VARIABLE SELECTION IN GENERALIZED LINEAR MIXED MODEL

Lanfeng Pan, Yehua Li

Department of Statistics & Statistical Laboratory, Iowa State University

Kevin He, Yanming Li and Yi Li

School of Public Health & Kidney Epidemiology and Cost Center, University of Michigan

Abstract

When the random effects of a Generalized Linear Mixed Model (GLMM) are drawn from a non-homogeneous population with subgroups but are assumed to follow a homogeneous Gaussian distribution, the predicted random effects are deceptively homogeneous. We propose a subgroup analysis approach on the random effects, allowing each level of random effect to have different means. The l_0 penalty on the pairwise difference of the individual means are used and the problem is solved by a greedy optimization algorithm similar to the hierarchical clustering. The random effects are automatically clustered into subgroups and the number of groups is determined by a tuning parameter. The tuning parameter is selected by cross-validation and minimizing the Mean Integrated Squared Error (MISE) of the proposed distribution for random effects. Simulations show the cross-validation and MISE method outperform other cluster selection methods. Our method can also accommodate high dimensional covariates. We impose a concave penalty on the fixed effects and perform variable selection on the fixed effects via BIC.

Key Words: Clustering; Health policy; Latent variables; Subgroup analysis; Variable selection.

3.1 Introduction

Our research is motivated from evaluating the performance of nationwide kidney transplant centers. Kidney transplantation, a primary therapy for end stage renal disease, is a complicated procedure typically involving transplant surgeons and physicians, coordinators, social workers, financial counselors, nutritionists, psychologists and referring physicians. The quality of care delivered by a transplant center is often assessed by patient survival, such as the 5 years post-transplant survival rate. A Generalized Linear Mixed Model (GLMM), considering both patient level risk factors and an effect representing the quality of care of the transplant centers, is often used to provide the assessment of each transplant center.

There are two different ways in treating the transplant center effects. Many statisticians and health policy researchers model the transplant center effects as random effects that follow a Gaussian distribution (Krumholz et al., 2006a,b; Li et al., 2009). Verbeke and Lesaffre (1996) show the random effects are badly estimated in linear mixed model with heterogeneity if normality is assumed. Agresti et al. (2004) show misspecifying the distribution of random effects reduces efficiencies. Komárek and Lesaffre (2008) show misspecification results in bias on fixed effects and the bias is larger for models with random effects of larger variance. This approach ignores the heterogeneity among the transplant centers and the assumption of a common Gaussian distribution induces a shrinkage effect that makes the predicted random effects similar in value. He et al. (2013) argue that borrowing information from other transplant centers is not fair when the goal of the study is to evaluate the centers and advocate modeling the transplant center effects as fixed effects. However, in such a fixed effects model, the number of parameters is large, making statistical inference numerically unstable, especially when the center size varies substantially.

Gallant and Ronald (1993); Chen et al. (2002); Ghidey et al. (2004) investigate into relaxing the Gaussian assumption on random effects. While it is interesting to model random effects with a flexible distribution and avoid misspecification, finding out the clusters in transplant center is a more important and challenging topic. In Chapter 3, the transplant

center effects are modeled as a Gaussian mixture distribution, which is essentially a model-based clustering. In model-based clustering, we need to supply the number of clusters before fitting the model, while the number of clusters needs to be decided via a sequential hypothesis testing approach, which does not consistently select the true model.

This chapter further investigates the problem under a more general clustering framework. Each transplant center is allowed to have an individual effect, which is equivalent to modeling the transplant center effects as fixed. By imposing a pairwise fusion penalty on the transplant center, transplant centers with effects close to each other will be merged into the same cluster. If the penalty is infinity all transplant center effects will be in one cluster with a common mean, resulting a Gaussian random effects model. If the random effects are heterogeneous or essentially come from several subgroups, we will choose a tuning parameter such that model is in the middle between a fixed effects model and a Gaussian distributed random effects model. By clustering transplant centers with similar performance together, we are able to find out the heterogeneity between underperforming or outperforming transplant centers. Furthermore by pooling information of multiple transplant centers from a cluster together, we reduce the number of parameters and obtain a stable estimation.

The clustering approach in this chapter is motivated from recent researches on convex clustering and subgroup analysis. Hocking et al. (2011) and Lindsten et al. (2011) formulate clustering with a convex penalty as a convex optimization problem. Chi and Lange (2015) propose to solve the convex clustering via Alternating Directions Method of Multipliers (Boyd et al., 2011) and Radchenko and Mukherjee (2017) find out the solution path to univariate convex clustering problem with a l_1 penalty. However, as shown by simulations of Wu et al. (2016) and Ma and Huang (2017), convex clustering detects no clusters and all observations merge into one cluster at the same time. Wu et al. (2016) and Ma and Huang (2017) propose to use concave penalties such as truncated lasso, SCAD (Fan and Li, 2001) and MCP (Zhang, 2010). Their algorithms can successfully detect clusters and give dendrograms clearly showing clusters. Motivated from their research, we find that the l_0 penalty can also achieve similar result while the computation time is dramatically reduced.

By imposing a concave penalty, variable selection on the fixed effects is also performed. The nonzero fixed effects are selected by fitting a model assuming all transplant centers having a common mean. Then the number of clusters are selected by cross-validation or minimizing the Mean Integrated Squared Error (MISE) of the proposed distribution while only the selected fixed effects are used. Our method is different from the random effects selection problem such as Ibrahim et al. (2011) and Hui et al. (2016). They are interested in selecting important variables among multiple random effect variables. Instead, we focus on multiple levels of one effect, i.e. the transplant centers.

The rest of this article is organized as follows. In section 2, the clustering algorithm is described in the situation when the transplant center effects are directly observed for ease of understanding. In section 3, the clustering algorithm on random effects is formally presented in the framework of GLMM. In section 4, the methods of selecting the number of clusters are described. In section 5, the proposed method is evaluated with simulations and in section 6, the proposed method is applied to the real data.

3.2 Subgroup Analysis when Transplant Center Effects Are Observed

We describe the subgroup analysis in the case when we directly observed the interested variable. Denote y_i and μ_i $(i = 1, \ldots, n)$ as the observations and their corresponding cluster means. The goal function is

$$\sum_{i=1}^{n}(y_i - \mu_i)^2 + \sum_{i_1<i_2} w_{i_1 i_2} p(|\mu_{i_1} - \mu_{i_2}|; \lambda) \tag{3.1}$$

where $w_{i_1 i_2}$ is a weight and $p(\cdot)$ is the penalty function. The weight $w_{i_1 i_2}$ is set to be 1 by most researchers (Radchenko and Mukherjee, 2017; Wu et al., 2016; Ma and Huang, 2017). Possible choices for the penalty functions including truncated lasso(Shen et al., 2012) , SCAD (Fan and Li, 2001) and MCP(Zhang, 2010).

We briefly review the penalty functions. Denote η as the difference between μs. The SCAD penalty (Fan and Li, 2001) is defined as

$$p(|\eta|;\lambda) = \begin{cases} \lambda|\eta| & if\ |\eta| \le \lambda; \\ \frac{2a\lambda|\eta|-\eta^2-\lambda^2}{2(a-1)} & if\ \lambda < |\eta| \le a\lambda; \\ \frac{\lambda^2(a+1)}{2} & if\ |\eta| > a\lambda, \end{cases}$$

where $a > 2$ is required. It is usually chosen to be 3.7 as recommended by Fan and Li (2001). The MCP penalty (Zhang, 2010) is

$$p(|\eta|;\lambda) = \begin{cases} \lambda|\eta| - \frac{\eta^2}{2a} & if\ |\eta| \le a\lambda; \\ \frac{1}{2}a\lambda^2 & if\ |\eta| > a\lambda, \end{cases}$$

where $a > 1$ is required and $a = 3$ is used in Ma and Huang (2017). The MCP is similar to SCAD in shape and also satisfies the oracle property. The truncated lasso penalty (Shen et al., 2012), aiming to be a close approximation to l_0 penalty, is defined as

$$p(|\eta|;\lambda) = \lambda \min(|\eta|, \tau),$$

where τ is another tuning parameter.

3.2.1 Subgroup analysis with l_0 penalty

Denote the clusters $\mathcal{G}_1, \ldots, \mathcal{G}_C$ as a non-overlapping split of the n observations and $|\mathcal{G}_c|$ as the group size. Denote $\mathcal{G}(i) = \mathcal{G}_c$ if $i \in \mathcal{G}_c$. In our algorithm, we choose $w_{i_1 i_2}$ as $\frac{1}{|\mathcal{G}(i_1)|+|\mathcal{G}(i_2)|}$ and use l_0 penalty. The motivation is given below

Consider the case when there are two groups, say $\mathcal{G}_1$ and $\mathcal{G}_2$ with means μ_1 and μ_2 respectively. Assume $w_{i_1 i_2} = 1$ for now. The goal function is

$$\sum_{i\in\mathcal{G}_1}(y_i-\mu_1)^2 + \sum_{i\in\mathcal{G}_2}(y_i-\mu_2)^2 + |\mathcal{G}_1||\mathcal{G}_2|p(|\mu_1-\mu_2|;\lambda).$$

If $\mathcal{G}_1$ and $\mathcal{G}_2$ are merged into a new group with $\mu_{new} = \frac{|\mathcal{G}_1|\mu_1+|\mathcal{G}_2|\mu_2}{|\mathcal{G}_1|+|\mathcal{G}_2|}$, the goal function will be

$$\begin{aligned} &\sum_{i\in\mathcal{G}_1}(y_i-\mu_{new})^2 + \sum_{i\in\mathcal{G}_2}(y_i-\mu_{new})^2 \\ &\qquad = \textstyle\sum_{i\in\mathcal{G}_1}(y_i-\mu_1)^2 + \sum_{i\in\mathcal{G}_2}(y_i-\mu_2)^2 + \frac{|\mathcal{G}_1||\mathcal{G}_2||\mu_1-\mu_2|^2}{|\mathcal{G}_1|+|\mathcal{G}_2|}. \end{aligned}$$

The difference is

$$\frac{|\mathcal{G}_1||\mathcal{G}_2||\mu_1-\mu_2|^2}{|\mathcal{G}_1|+|\mathcal{G}_2|}-|\mathcal{G}_1||\mathcal{G}_2|p(|\mu_1-\mu_2|;\lambda). \tag{3.2}$$

It is easy to see the penalty function is dominating as the cluster size $|\mathcal{G}_1|+|\mathcal{G}_2|$ increase, which can be as large as n. The concave penalty term will work like a hard thresholding since the first order derivative of the penalty term is exploding as $|\mathcal{G}_1|+|\mathcal{G}_2|$ increases. The difference $|\mu_1-\mu_2|$ will be shrunk to 0 if $|\mu_1-\mu_2|<O(\frac{1}{|\mathcal{G}_1|+|\mathcal{G}_2|})$. Otherwise, the penalty term will be flat and $|\mu_1-\mu_2|$ are not shrunk at all.

Motivated by the fact that the behavior of the algorithms in Ma and Huang (2017) and Wu et al. (2016) are similar to the l_0 penalty, we decide to use l_0 penalty in our algorithm. We find our algorithm gives similar results as in Ma and Huang (2017) and Wu et al. (2016) while our algorithm requires less computation. In this article, we use the following penalty

$$p(|\mu_{i_1}-\mu_{i_2}|;\lambda)=\lambda^2 I(\mu_{i_1}\neq\mu_{i_2}).$$

With this l_0 penalty and our choice of $w_{i_1i_2}=\frac{1}{|\mathcal{G}(i_1)|+|\mathcal{G}(i_2)|}$, our goal function becomes

$$\sum_{i=1}^{n}(y_i-\mu_i)^2+\sum_{i_1<i_2}\frac{\lambda^2 I(\mu_{i_1}\neq\mu_{i_2})}{|\mathcal{G}(i_1)|+|\mathcal{G}(i_2)|}. \tag{3.3}$$

The difference (3.2) becomes

$$\frac{|\mathcal{G}_1||\mathcal{G}_2||\mu_1-\mu_2|^2}{|\mathcal{G}_1|+|\mathcal{G}_2|}-\frac{|\mathcal{G}_1||\mathcal{G}_2|}{|\mathcal{G}_1|+|\mathcal{G}_2|}\lambda^2.$$

Clearly, if $\lambda>|\mu_1-\mu_2|$, the goal function will be minimized if $\mathcal{G}_1$ and $\mathcal{G}_2$ merge. If $\lambda<|\mu_1-\mu_2|$, then $\mathcal{G}_1$ and $\mathcal{G}_2$ should stay separate.

Remark 3 (3.3) is different when $\mathcal{G}_c$ for $c=1,\ldots,C$ are different. Thus, the estimation depends on the current group assignments. Given the current grouping status and λ, we can find the optimal solution for (3.3).

Our algorithm is similar to agglomerative hierarchical clustering (Ward, 1963; Friedman et al., 2001). We briefly describe the algorithm of agglomerative hierarchical clustering with group average link in following. Denote the n observations as y_i $(i=1,\ldots,n)$ and their

dissimilarity is measured by $(y_i - y_j)^2$ for univariate case. Other dissimilarity can be defined but we focus on the simple case. Given a dissimilarity $d(\cdot, \cdot)$ between clusters, the algorithm is as following.

Step 1 Starting with $s = 0$, $\lambda^{(0)} = 0$, $\mathcal{G}_i = \{y_i\}$ *and* $\mu_i^{(0)} = y_i$ *for* $i = 1, \ldots, C$ *where* $C = n$.

Step 2 Update $s \leftarrow s + 1$. *Update* $\eta_{c_1 c_2} = d(\mathcal{G}_{c_1}, \mathcal{G}_{c_2})$ *for* $1 \le c_1 < c_2 \le C$.

Step 3 Find out the smallest $|\eta_{c_1 c_2}|$ *and set* $\lambda^{(s)} = |\eta_{c_1 c_2}|$. *Merge* $\mathcal{G}_{c_1} \leftarrow \mathcal{G}_{c_1} \cup \mathcal{G}_{c_2}$. *Remove* $\mathcal{G}_{c_2}$ *and rename* $\mathcal{G}_{c-1} \leftarrow \mathcal{G}_c$ *for* $c_2 < c \le n$. *Update* $C \leftarrow C - 1$

Step 4 Update $\mu(\mathcal{G}_c) = \sum_{i \in \mathcal{G}_c} y_i / |\mathcal{G}_c|$ *and* $\mu_i^{(s)} = \mu(\mathcal{G}_c)$ *if* $i \in \mathcal{G}_c$ *for* $i = 1, \ldots, C$.

Step 5 Go to Step 2 if $C > 1$. *Otherwise stop and output* $\lambda^{(s)}$ *and* $\{\mu_i^{(s)}\}_{i=1}^n$ *for* $s = 0, 1 \ldots$.

There are many different definitions for the cluster dissimilarity d. Some common choices are single linkage or complete linkage. Here we use the group average linkage

$$d(\mathcal{G}_{c_1}, \mathcal{G}_{c_2}) = \frac{\sum_{i \in \mathcal{G}_{c_1}, j \in \mathcal{G}_{c_2}} (y_i - y_j)^2}{|\mathcal{G}_{c_1}||\mathcal{G}_{c_2}|}.$$

If we define cluster mean as $\mu(\mathcal{G}_c) = \sum_{i \in \mathcal{G}_c} y_i / |\mathcal{G}_c|$, then the between group average link $d(\mathcal{G}_{c_1}, \mathcal{G}_{c_2}) = \{\mu(\mathcal{G}_{c_1}) - \mu(\mathcal{G}_{c_2})\}^2$. See Murtagh (1983), Podani (1989) and Friedman et al. (2001) for more details.

Our algorithm is very similar to hierarchical clustering with some modifications. Denote $y_{max} = \max\{y_i, i = 1, \ldots, n\}$, $y_{min} = \min\{y_i, i = 1, \ldots, n\}$ and $R = y_{max} - y_{min}$.

Step 1 Starting with $s = 0$, $\lambda^{(1)} = \frac{R}{nlog(n)}$, $\mu_i^{(0)} = y_i$, *and* $\mathcal{G}_i = \{y_i\}$ *for* $i = 1, \ldots, C$ *where* $C = n$.

Step 2 Update $\eta_{c_1 c_2} = \mu(\mathcal{G}_{c_1}) - \mu(\mathcal{G}_{c_2})$ *for* $1 \le c_1 < c_2 \le C$.

Step 3 Set the new $\mathcal{G}_{c_1} \leftarrow \mathcal{G}_{c_1} \cup \mathcal{G}_{c_2}$ *if* $|\eta_{c_1 c_2}| \le \lambda^{(s)}$. *Remove* $\mathcal{G}_{c_2}$ *and rename* $\mathcal{G}_{c-1} \leftarrow \mathcal{G}_c$ *for* $c_2 \le c \le n$. *Update* $C \leftarrow C - 1$.

Step 4 Update $\mu(\mathcal{G}_c) = \sum_{i_1 \in \mathcal{G}_c} y_{i_1} / |\mathcal{G}_c|$ *for* $c = 1, \ldots, C$ *and* $\mu_i^{(s)} = \mu(\mathcal{G}_c)$ *if* $i \in \mathcal{G}_c$ *for* $i = 1, \ldots, n$.

Step 5 Update $s \leftarrow s+1$. *Set* $\lambda^{(s)}$ *as the smallest nonzero values in* $|\eta_{c_1c_2}|$ *for* $1 \le c_1 < c_2 \le C$. *Go to Step 2 if* $C > 1$. *Otherwise stop and output* $\lambda^{(s)}$ *and* $\{\mu_i^{(s)}\}_{i=1}^n$ *for* $s = 0, 1 \ldots$.

At step 1, we set $\lambda^{(1)} = \frac{R}{n\log(n)}$ instead of the smallest nonzero values in $|\eta_{i_1 i_2}|$ for $1 \le i_1 < i_2 \le n$. By choosing a large $\lambda^{(1)}$, we save many early stage iterations which is usually neither important nor interesting. The value $\frac{R}{n\log(n)}$ can approximately reduce the current n clusters by half. We can further save computation times by choosing a larger $\lambda^{(1)}$. Our simulation show $\lambda^{(1)} = \frac{R}{n\log(n)}$ works well. The computation complexity of our algorithm is $O(n^2)$, the same as hierarchical clustering. In contrast the algorithm in Ma and Huang (2017) has the complexity of $O(n^3)$ because they need to inverse a matrix of n by n at each iteration. The algorithm in Wu et al. (2016) has complexity of $O(n^2/\epsilon)$ each iteration where ϵ is a small tolerance constant. Wu et al. (2016)'s algorithm does not allow additional covariates and are not applicable to GLMM. To compare the speed, we generate a dataset with 300 observations. Our algorithm takes 0.68s to find out the complete dendrogram while the algorithm of Ma and Huang (2017) needs 21.03s. Both algorithms are implemented in Julia language and are tested on a personal laptop.

3.3 Subgroup analysis and variable selection in GLMM

Now we describe our model under the framework of GLMM. Suppose that there are n independent transplant centers, each treating N_i patients, which brings the total sample size to be $N = \sum_{i=1}^n N_i$. Let Y_{ik} be the outcome variable of the kth patient treated at the ith transplant center and let $\boldsymbol{X}_{ik} \in \mathbb{R}^J$ be the patient level covariate for $k = 1, \ldots N_i$ and $i = 1, \ldots n$. Denote $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iN_i})^{\mathrm{T}}$ and $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iN_i})^{\mathrm{T}}$. Let γ_i be the random effect that represents the care qualify of the ith center and denote $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^T$. The random effects $\boldsymbol{\gamma}$ are unobserved. Suppose the conditional density of Y_{ik} given γ_i belongs to the canonical exponential family:

$$f(Y_{ik}|\boldsymbol{X}_{ik}, \gamma_i; \boldsymbol{\beta}, \varphi) = \exp\left\{\frac{Y_{ik}\xi_{ik} + b(\xi_{ik})}{a(\varphi)} + d(Y_{ik}, \varphi)\right\}, \tag{3.4}$$

where $a(\cdot)$, $b(\cdot)$ and $d(\cdot)$ are known functions, $\xi_{ik} = \boldsymbol{X}_{ik}^T\boldsymbol{\beta} + \gamma_i$ is the canonical parameter with $\mathrm{E}(Y_{ik}|\boldsymbol{X}_{ik}, \gamma_i) = b'(\xi_{ik})$, and φ is a nuisance parameter.

Assume that Y_{ik} and $Y_{ik'}$ are independent given γ_i for any $k \neq k'$. In our transplant center evaluation application, we consider a binary response variable: $Y_{ik} = 1$ if the patient is deceased within 5 years after transplant; -1 otherwise. In the dataset, there were essentially no censoring within the first 5 years as the transplant patients' survival information had been closely monitored and tracked. This gives the justification of treating 5 year survival as a binary outcome data. With that, model (3.4) becomes $f(Y_{ik}|\boldsymbol{X}_{ik}, \gamma_i; \boldsymbol{\beta}) = \{1 + \exp(-\xi_{ik} Y_{ik})\}^{-1}$. The effect of transplant center i is described by a Gaussian distribution with mean μ_i $(i = 1, \ldots, n)$ and variance σ^2. For convenience we denote $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ and $\boldsymbol{\theta}_\gamma = (\mu_1, \ldots, \mu_n, \sigma)^T$. Let $\mathcal{G}_1, \ldots, \mathcal{G}_C$ be a split of all transplant centers and all the transplant centers in one group have equal means. That is $\mu_i = \mu(\mathcal{G}_c)$ for all $i \in \mathcal{G}_c$ where $\mu(\mathcal{G}_c)$ denotes the common mean for group $\mathcal{G}_c$. At first, each transplant center is allowed to have an individual mean, i.e. $\mathcal{G}_i = \{i\}$ for $i = 1, \ldots, n$ and $C = n$.

Denote $\boldsymbol{Y} = (\boldsymbol{Y}_1^{\mathrm{T}}, \ldots, \boldsymbol{Y}_n^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{X} = (\boldsymbol{X}_1^{\mathrm{T}}, \ldots, \boldsymbol{X}_n^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\theta} = (\boldsymbol{\mu}^T, \sigma, \boldsymbol{\beta}^T)^T$. Then the likelihood function for the complete data, comprising of both observed and latent variables, is

$$l_{comp}(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \ell_{i,comp}(\boldsymbol{\theta}; \boldsymbol{Y}_i, \boldsymbol{X}_i, \gamma_i),$$

where

$$\ell_{i,comp}(\boldsymbol{\theta}; \boldsymbol{Y}_i, \boldsymbol{X}_i, \gamma_i) = \log f(\boldsymbol{Y}_i|\boldsymbol{X}_i, \gamma_i; \boldsymbol{\beta}) - \frac{1}{2}\log(\sigma^2) + \log\phi\{(\gamma_i - \mu_i)/\sigma\}$$

and $f(\boldsymbol{Y}_i|\boldsymbol{X}_i, \gamma_i; \boldsymbol{\beta}) = \prod_{k=1}^{N_i} f(Y_{ik}|\boldsymbol{X}_{ik}, \gamma_i; \boldsymbol{\beta})$. Thus, the log complete likelihood is

$$\begin{aligned} l_{comp}(\boldsymbol{\theta}) \quad \propto \quad & \sum_{i=1}^{n}\sum_{k=1}^{N_i} \log f(Y_{ik}|\boldsymbol{X}_{ik}, \gamma_i; \boldsymbol{\beta}) - \sum_{i=1}^{n} \frac{(\gamma_i - \mu_i)^2}{2\sigma^2} \\ & -\frac{1}{2}\sum_{i=1}^{n} \log(\sigma^2). \end{aligned} \tag{3.5}$$

A penalty on the fixed effects $\boldsymbol{\beta}$

$$p(\boldsymbol{\beta}; \lambda_\beta) = -N \sum_{j=1}^{J} p(|\beta|_j; \lambda_\beta)$$

is used for variable selection purpose. The p function can be chosen from lasso, SCAD or MCP. The penalized complete log likelihood is

$$pl_{comp}(\boldsymbol{\theta}) = l_{comp}(\boldsymbol{\theta}) + p(\boldsymbol{\beta}; \lambda_\beta) - \sum_{i_1 < i_2} \frac{\lambda_\gamma^2 I(\mu_{i_1} \neq \mu_{i_2})}{2\sigma^2(|\mathcal{G}(i_1)| + |\mathcal{G}(i_2)|)}.$$

Integrating the complete likelihood over $\boldsymbol{\gamma}$, we obtain the log marginal likelihood

$$l_{mar}(\boldsymbol{\theta};\boldsymbol{Y},\boldsymbol{X}) = \sum_{i=1}^{n} l_{i,mar}(\boldsymbol{\theta};\boldsymbol{Y}_i,\boldsymbol{X}_i) \tag{3.6}$$

where

$$l_{i,mar}(\boldsymbol{\theta}) = \log \int \prod_{k=1}^{N_i} f(Y_{ik}|\boldsymbol{X}_{ik},\gamma_i;\boldsymbol{\beta}) f(\gamma|\mu_i,\sigma) d\gamma_i.$$

The complete log likelihood cannot be maximized easily since $\boldsymbol{\gamma}$ is latent and unknown. Denote $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t-1)})$ as the expectation of $\ell_{comp}(\boldsymbol{\theta})$ on posterior distribution of γ given observed data and current estimation $\widehat{\boldsymbol{\theta}}^{(t-1)}$. The Expectation-Maximization (EM) algorithm maximizes $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t-1)})$ and $\widehat{\boldsymbol{\theta}}^{(t)} = \arg\max Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t-1)})$. Specifically

$$\begin{aligned} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t-1)}) &= \sum_{i=1}^{n} \int \log f(\boldsymbol{Y}_i|\boldsymbol{X}_i,\gamma;\boldsymbol{\beta}) f(\gamma|\boldsymbol{X}_i,\boldsymbol{Y}_i;\widehat{\boldsymbol{\theta}}^{(t-1)}) d\gamma \\ &\quad + \sum_{i=1}^{n} \int \left[-\frac{(\gamma_i-\mu_i)^2}{2\sigma^2} - \frac{1}{2}\log(\sigma^2) \right] f(\gamma|\boldsymbol{X}_i,\boldsymbol{Y}_i;\widehat{\boldsymbol{\theta}}^{(t-1)}) d\gamma \\ &\quad + p(\boldsymbol{\beta};\lambda_\beta) - \sum_{i_1<i_2} \frac{I(\mu_{i_1}\neq\mu_{i_2})\lambda_\gamma^2}{2\sigma^2(|\mathcal{G}(i_1)|+|\mathcal{G}(i_2)|)} \end{aligned}$$

where

$$f(\gamma|\boldsymbol{X}_i,\boldsymbol{Y}_i;\widehat{\boldsymbol{\theta}}^{(t-1)}) = \frac{f(\boldsymbol{Y}_i|\boldsymbol{X}_i,\gamma;\widehat{\boldsymbol{\beta}}^{(t-1)}) \frac{1}{\widehat{\sigma}^{(t-1)}} \phi\left(\frac{\gamma-\widehat{\mu}_i^{(t-1)}}{\widehat{\sigma}_i^{(t-1)}}\right)}{\int f(\boldsymbol{Y}_i|\boldsymbol{X}_i,\gamma;\widehat{\boldsymbol{\beta}}^{(t-1)}) \frac{1}{\widehat{\sigma}^{(t-1)}} \phi\left(\frac{\gamma-\widehat{\mu}_i^{(t-1)}}{\widehat{\sigma}^{(t-1)}}\right) d\gamma}$$

is the posterior density of γ_i conditioning on data $\boldsymbol{X}_i$, $\boldsymbol{Y}_i$ and $\widehat{\boldsymbol{\theta}}^{(t-1)}$.

Define

$$Q_1(\boldsymbol{\beta}|\widehat{\boldsymbol{\theta}}^{(t-1)}) = \sum_{i=1}^{n} \int \log f(\boldsymbol{Y}_i|\boldsymbol{X}_i,\gamma;\boldsymbol{\beta}) f(\gamma|\boldsymbol{X}_i,\boldsymbol{Y}_i;\widehat{\boldsymbol{\theta}}^{(t-1)}) d\gamma + p(\boldsymbol{\beta};\lambda_\beta)$$

and

$$\begin{aligned} Q_2(\boldsymbol{\theta}_\gamma|\widehat{\boldsymbol{\theta}}^{(t-1)}) &= \sum_{i=1}^{n} \int \left[-\frac{(\gamma_i-\mu_i)^2}{2\sigma^2} - \frac{1}{2}\log(\sigma^2) \right] f(\gamma|\boldsymbol{X}_i,\boldsymbol{Y}_i;\widehat{\boldsymbol{\theta}}^{(t-1)}) d\gamma \\ &\quad - \sum_{i_1<i_2} \frac{I(\mu_{i_1}\neq\mu_{i_2})\lambda_\gamma^2}{2\sigma^2(|\mathcal{G}(i_1)|+|\mathcal{G}(i_2)|)} \end{aligned}$$

then

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t-1)}) = Q_1(\boldsymbol{\beta}|\widehat{\boldsymbol{\theta}}^{(t-1)}) + Q_2(\boldsymbol{\theta}_\gamma|\widehat{\boldsymbol{\theta}}^{(t-1)}). \tag{3.7}$$

Clearly, Q_1 only depends on fixed effects $\boldsymbol{\beta}$ and Q_2 only depends on random effects parameters $\boldsymbol{\theta}_\gamma$. The two parts Q_1 and Q_2 can be maximized separately in each iteration. The integrals in Q_1 and Q_2 are evaluated via Gauss-Hermite approximation.

We describe our algorithm for fixed λ_β below.

Step 1 Initializing $\widehat{\boldsymbol{\theta}}^{(0)}$ *and* $\{\gamma_i^{(0)}\}_{i=1}^n$. *Let* $R = \max\{\gamma_i^{(0)}\}_{i=1}^n - \min\{\gamma_i^{(0)}\}_{i=1}^n$. *Setting* $s = 1$ *and* $\lambda_\gamma^{(1)} = \frac{R}{nlogn}$

Step 2 Set $t = 1$. *Repeat the following procedure until convergence of* $\widehat{\boldsymbol{\theta}}^{(t)}$*:*

1. *Evaluate* $Q_1(\boldsymbol{\beta}|\widehat{\boldsymbol{\theta}}^{(t-1)})$ *and update* $\boldsymbol{\beta}^{(t)}$ *for fixed* λ_β.
2. *Evaluate* $Q_2(\boldsymbol{\theta}_\gamma|\widehat{\boldsymbol{\theta}}^{(t-1)})$ *and update* $\widehat{\mu}_i^{(t)}$ *and* $\widehat{\sigma}^{(t)}$ *at* $\lambda_\gamma^{(s)}$.
3. $t \leftarrow t + 1$.

Step 3 Record current $\widehat{\boldsymbol{\theta}}$ *and* $\lambda_\gamma^{(s)}$. *Update* $\eta_{i_1 i_2} = \mu_{i_1} - \mu_{i_2}$ *for* $1 \le i_1 < i_2 \le n$. *Stop if all* $|\eta_{i_1 i_2}|$ *are 0. Otherwise set* $s \leftarrow s + 1$ *and* $\lambda_\gamma^{(s)}$ *as the smallest nonzero value of* $|\eta_{i_1 i_2}|$ *for* $1 \le i_1 < i_2 \le n$. *Go to step 2.*

Details about Gauss-Hermite approximation, initialization and stopping criteria, updating $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\theta}}_\gamma$ are given the following subsections.

3.3.1 Gauss-Hermite Approximation

To deal with the integrals in Q, we use the Gauss-Hermite approximation. Integrals with respect to a Gaussian density can be well approximated by Gauss-Hermite quadrature:

$$\int h(\gamma)\frac{1}{\sigma}\phi\{(\gamma-\mu)/\sigma\}d\gamma \approx \frac{1}{\sqrt{\pi}}\sum_{m=1}^{M} v_m h(\gamma^{(m)})$$

where $h(\gamma)$ is an integrable real valued function, $\gamma^{(m)} = \mu + \sqrt{2}\sigma d_m$, $d_1, d_2, \ldots, d_M$ are the Gauss-Hermite abscissas and $v_1, \ldots, v_M$ are the corresponding quadrature weights. We find in our numerical studies that using $M = 100$ quadrature points usually provides a close enough approximation.

Then Q_1 can be well approximated by

$$\sum_{i=1}^{n}\sum_{m=1}^{M}\omega_i^{(m,t-1)}\log f(\boldsymbol{Y}_i|\boldsymbol{X}_i,\gamma_i^{(m,t-1)};\boldsymbol{\beta})+p(\boldsymbol{\beta};\lambda_\beta)$$

and Q_2 by

$$-\sum_{i=1}^{n}\sum_{m=1}^{M}\omega_i^{(m,t-1)}\frac{(\gamma_i^{(m,t-1)}-\mu_i)^2}{2\sigma^2}-\frac{n}{2}\log(\sigma^2)$$

where $\omega_i^{(m,t-1)}$ is defined as

$$\omega_i^{(m,t-1)}=\frac{v_m f(\boldsymbol{Y}_i|\boldsymbol{X}_i,\gamma_i^{(m,t-1)};\widehat{\boldsymbol{\beta}}^{(t-1)})}{\sum_{m=1}^{M}v_m f(\boldsymbol{Y}_i|\boldsymbol{X}_i,\gamma_i^{(m,t-1)};\widehat{\boldsymbol{\beta}}^{(t-1)})}. \tag{3.8}$$

The evaluating points $\gamma_i^{(m,t-1)}$ is the linear transformation of Gauss-Hermite abscissas d_m for $m=1,\ldots,M$.

$$\gamma_i^{(m,t-1)}=\widehat{\mu}_i^{(t-1)}+\sqrt{2}\widehat{\sigma}^{(t-1)}d_m.$$

3.3.2 Updating fixed effects

The coefficient $\widehat{\boldsymbol{\beta}}^{(t)}$ is obtained by maximizing

$$\sum_{i=1}^{n}\sum_{k=1}^{N_i}\sum_{m=1}^{M}\omega_i^{(m,t-1)}\log f(Y_{ik}|\boldsymbol{X}_{ik},\gamma_i^{(m,t-1)};\boldsymbol{\beta})-N\sum_{j=1}^{J}p(\mid\beta_j\mid;\lambda_\beta) \tag{3.9}$$

using the Alternating Direction Method of Multipliers (Boyd et al., 2011, ADMM). The ADMM reformulates (3.9) to be

$$\sum_{i=1}^{n}\sum_{k=1}^{N_i}\sum_{m=1}^{M}\omega_i^{(m,t-1)}\log f(Y_{ik}|\boldsymbol{X}_{ik},\gamma_i^{(m,t-1)};\boldsymbol{\beta})-N\sum_{j=1}^{J}p(\mid\eta_{\beta j}\mid;\lambda_\beta)$$

$$\text{subject to } \beta_j=\eta_{\beta j},\ j=1,\ldots,J,$$

where $\boldsymbol{\eta}_\beta=(\eta_{\beta,1},\ldots,\eta_{\beta,J})^T$ are intermediate variables. Then ADMM maximizes the following augmented Lagrangian

$$\begin{aligned}
Q_{aug}(\boldsymbol{\beta},\boldsymbol{\eta}_\beta,\boldsymbol{\upsilon}_\beta) &= \sum_{i=1}^{n}\sum_{k=1}^{N_i}\sum_{m=1}^{M}\omega_i^{(m,t-1)}\log f(Y_{ik}|\boldsymbol{X}_{ik},\gamma_i^{(m,t-1)};\boldsymbol{\beta})-N\sum_{j=1}^{J}\upsilon_{\beta j}(\beta_j-\eta_{\beta j})\\
&\quad -N\frac{\rho}{2}\sum_{j=1}^{J}(\beta_j-\eta_{\beta j})^2-N\sum_{j=1}^{J}p(|\eta_{\beta j}|;\lambda_\beta)\\
&= \sum_{i=1}^{n}\sum_{k=1}^{N_i}\sum_{m=1}^{M}\omega_i^{(m,t-1)}\log f(Y_{ik}|\boldsymbol{X}_{ik},\gamma_i^{(m,t-1)};\boldsymbol{\beta})\\
&\quad -N\frac{\rho}{2}\sum_{j=1}^{J}\left(\beta_j-\eta_{\beta j}+\frac{\upsilon_{\beta j}}{\rho}\right)^2+N\sum_{j=1}^{J}\frac{\upsilon_{\beta j}^2}{2\rho}-N\sum_{j=1}^{J}p(|\eta_{\beta j}|;\lambda_\beta)
\end{aligned}$$

where $\boldsymbol{\upsilon}_\beta = (\upsilon_{\beta 1}, \ldots, \upsilon_{\beta J})^T$ are Lagrangian multipliers and ρ is the ADMM penalty parameter which is set to be 1 throughout this paper.

We maximize $Q_{aug}(\boldsymbol{\beta}, \boldsymbol{\eta}_\beta, \boldsymbol{\upsilon}_\beta)$ over $\boldsymbol{\beta}$, then $\boldsymbol{\eta}_\beta$ and lastly $\boldsymbol{\upsilon}_\beta$. The part that related to $\boldsymbol{\beta}$ is

$$Q_{aug,\beta}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{k=1}^{N_i}\sum_{m=1}^{M} \omega_i^{(m,t-1)} \log f(Y_{ik}|\boldsymbol{X}_{ik}, \gamma_i^{(m,t-1)}; \boldsymbol{\beta}) \\ -N\tfrac{\rho}{2}\textstyle\sum_{j=1}^{J}\left(\beta_j - \eta_{\beta j}^{(t-1)} + \frac{\upsilon_{\beta j}^{(t-1)}}{\rho}\right)^2. \tag{3.10}$$

The maximization can be found by Newton-Raphson algorithm in general. In the case of binary response, the Iteratively Reweighted Least Squares (IRLS) can be used. Denote $p_{ik}^{(m,t-1,r)} = E(Y_{ik}|X_{ik}, \gamma_i^{(m,t-1)}; \widehat{\boldsymbol{\beta}}^{(t-1,r)})$ where $\widehat{\boldsymbol{\beta}}^{(t-1,r)}$ is the estimation at rth IRLS iteration. The quadratic approximation to $\log f(Y_{ik}|\boldsymbol{X}_{ik}, \gamma_i^{(m,t-1)}; \boldsymbol{\beta})$ evaluated at the $\widehat{\boldsymbol{\beta}}^{(t-1,r)}$ is

$$-\frac{1}{2}p_{ik}^{(m,t-1,r)}\left\{1 - p_{ik}^{(m,t-1,r)}\right\}\left[\boldsymbol{X}_{ik}\boldsymbol{\beta} - \boldsymbol{X}_{ik}\widehat{\boldsymbol{\beta}}^{(t-1,r)} - \frac{Y_{ik} - p_{ik}^{(m,t-1,r)}}{p_{ik}^{(m,t-1,r)}\left\{1 - p_{ik}^{(m,t-1,r)}\right\}}\right]^2,$$

with some term free from $\boldsymbol{\beta}$ omitted. Then $\boldsymbol{\beta}$ is updated by

$$\widehat{\boldsymbol{\beta}}^{(t,r)} = \widehat{\boldsymbol{\beta}}^{(t,r-1)} + q \times \left[\sum_{i=1}^{n}\sum_{k=1}^{N_i}\sum_{m=1}^{M} \omega_i^{(m,t-1)} p_{ik}^{(m,t-1,r)}\left\{1 - p_{ik}^{(m,t-1,r)}\right\}\boldsymbol{X}_{ik}\boldsymbol{X}_{ik}^T + \rho N I_J\right]^{-1} \\ \left[\sum_{i=1}^{n}\sum_{k=1}^{N_i}\sum_{m=1}^{M} \omega_i^{(m,t-1)}\left\{Y_{ik} - p_{ik}^{(m,t-1,r)}\right\}\boldsymbol{X}_{ik}^T - N\rho\left\{\widehat{\boldsymbol{\beta}}^{(t,r-1)} - \eta_{\beta j}^{(t-1)} + \frac{\upsilon_{\beta j}^{(t-1)}}{\rho}\right\}\right]$$

until

$$\frac{Q_{aug,\beta}(\widehat{\boldsymbol{\beta}}^{(t,r)}) - Q_{aug,\beta}(\widehat{\boldsymbol{\beta}}^{(t,r-1)})}{\mid Q_{aug,\beta}(\widehat{\boldsymbol{\beta}}^{(t,r-1)}) \mid} < 0.001.$$

The step size q is chosen to guarantee (3.10) is increased by $\widehat{\boldsymbol{\beta}}^{(t,r)}$ and I_J is a $J \times J$ identify matrix. If r_{cvg} iterations is used for IRLS in total, set $\widehat{\boldsymbol{\beta}}^{(t)} = \widehat{\boldsymbol{\beta}}^{(t,r_{cvg})}$.

After obtaining $\widehat{\boldsymbol{\beta}}^{(t)}$, $\boldsymbol{\eta}_\beta$ is updated by maximizing

$$Q_{aug,\eta}(\boldsymbol{\eta}_\beta) = -N\frac{\rho}{2}\sum_{j=1}^{J}\left(\widehat{\beta}_j^{(t)} + \frac{\upsilon_{\beta j}^{(t-1)}}{\rho} - \eta_{\beta j}\right)^2 - N\sum_{j=1}^{J} p(|\eta_{\beta j}|; \lambda_\beta).$$

Depending on the penalty function of $p(\cdot)$, $\eta_{\beta j}$ is updated differently. Denote $\delta_{\beta j}^{(t)} = \widehat{\beta}_j^{(t)} + \upsilon_{\beta j}^{(t-1)}/\rho$. In the case when SCAD penalty is used, the solution is given by

$$\eta_{\beta j}^{(t)} = \begin{cases} ST(\delta_{\beta j}^{(t)}, \lambda_\beta/\rho) & if\ \lambda_\beta/\rho < \delta_{\beta j}^{(t)} \leq \lambda_\beta/\rho + \lambda_\beta; \\ \frac{ST(\delta_{\beta j}^{(t)}, 1.37\lambda_\beta/\rho)}{1-0.37/\rho} & if\ \lambda_\beta/\rho + \lambda_\beta < \delta_{\beta j}^{(t)} \leq 3.7\lambda_\beta; \\ \delta_{\beta j}^{(t)} & if\ \delta_{\beta j}^{(t)} > 3.7\lambda_\beta, \end{cases}$$

where $ST(\delta_{\beta j}^{(t)}, \lambda_\beta/\rho)$ is the soft thresholding rule and

$$ST(\delta_{\beta j}^{(t)}, \lambda_\beta/\rho) = \begin{cases} 0 & if\ \mid \delta_{\beta j}^{(t)} \mid \leq \lambda_\beta/\rho; \\ \delta_{\beta j}^{(t)} \left(1 - \left|\frac{\lambda_\beta}{\rho \delta_{\beta j}^{(t)}}\right|\right) & otherwise. \end{cases}$$

Lastly, $\boldsymbol{\upsilon}_\beta$ is updated by

$$\upsilon_{\beta j}^{(t)} = \upsilon_{\beta j}^{(t-1)} + \rho(\widehat{\beta}_j^{(t)} - \eta_{\beta j}^{(t)})$$

for $j = 1, \ldots, J$.

3.3.3 Clustering the random effects

The random effects parameters $\boldsymbol{\mu}$ and σ^2 only depends on Q_2.

$$-\frac{\sum_{i=1}^{n}\sum_{m=1}^{M}\omega_i^{(m,t-1)}(\gamma_i^{(m,t-1)} - \mu_i)^2}{2\sigma^2} - \sum_{i_1<i_2}\frac{I(\mu_{i_1} \neq \mu_{i_2})\lambda_\gamma^2}{2\sigma^2(|\mathcal{G}(i_1)| + |\mathcal{G}(i_2)|)} - \frac{n}{2}\log(\sigma^2).$$

For convenience, denote

$$L(\boldsymbol{\mu}; \lambda_\gamma) = \sum_{i=1}^{n}\sum_{m=1}^{M}\omega_i^{(m,t-1)}(\gamma_i^{(m,t-1)} - \mu_i)^2 + \sum_{i_1<i_2}\frac{I(\mu_{i_1} \neq \mu_{i_2})\lambda_\gamma^2}{|\mathcal{G}(i_1)| + |\mathcal{G}(i_2)|}. \tag{3.11}$$

For any $\boldsymbol{\mu}(\boldsymbol{\mu})$, the σ^2 that maximize Q_2 is

$$\sigma^2(\boldsymbol{\mu}) = \frac{1}{n}L(\boldsymbol{\mu}; \lambda_\gamma).$$

Profiling on σ^2, $\boldsymbol{\mu}$ is updated by maximizing

$$-\frac{n}{2}\log\{L(\boldsymbol{\mu}; \lambda_\gamma)\} - \frac{n}{2} + \frac{n}{2}\log(n)$$

or minimizing $L(\boldsymbol{\mu}; \lambda_\gamma)$ to be more specific. The $L(\boldsymbol{\mu}; \lambda_\gamma)$ is similar to the goal function defined in 3.3. Thus, $\boldsymbol{\mu}$ can be updated similarly.

Define

$$\widehat{\mu}(\mathcal{G}_c) = \frac{\sum_{i\in\mathcal{G}_c}\widehat{\gamma}_i^{(t)}}{|\mathcal{G}_c|}$$

for $i \in \mathcal{G}_c$ and $c = 1, \ldots, C$ where

$$\widehat{\gamma}_i^{(t)} = \sum_{m=1}^{M} \omega_i^{(m,t-1)}\gamma_i^{(m,t-1)}$$

is the posterior mean of γ_i conditioning on current estimate $\widehat{\boldsymbol{\theta}}^{(t-1)}$.

If $|\widehat{\mu}(\mathcal{G}_{c_1}) - \widehat{\mu}(\mathcal{G}_{c_2})| > \lambda_\gamma$ for $1 \le c_1 < c_2 \le C$, no merging is needed. The $\boldsymbol{\mu}$ is updated by

$$\widehat{\mu}_i^{(t)} = \widehat{\mu}(\mathcal{G}_c)$$

for $i \in \mathcal{G}_c$ and $c = 1, \ldots, C$. The current group assignments $\mathcal{G}_c$ (c=1,...,C) remain unchanged. Denote the result of 3.11 as L_C for convenience.

If $|\widehat{\mu}(\mathcal{G}_{c_1}) - \widehat{\mu}(\mathcal{G}_{c_2})| \le \lambda_\gamma$, merge $\mathcal{G}_{c_1}$ and $\mathcal{G}_{c_2}$ into a new group $\mathcal{G}_{new}$ with group mean

$$\widehat{\mu}_{new} = \frac{\sum_{i\in\mathcal{G}_{c_1}\cup\mathcal{G}_{c_2}}\widehat{\gamma}_i^{(t)}}{|\mathcal{G}_{c_1}| + |\mathcal{G}_{c_2}|} = \frac{|\mathcal{G}_{c_1}|\widehat{\mu}(\mathcal{G}_{c_1}) + |\mathcal{G}_{c_2}|\widehat{\mu}(\mathcal{G}_{c_2})}{|\mathcal{G}_{c_1}| + |\mathcal{G}_{c_2}|}.$$

Update $\widehat{\mu}_i = \widehat{\mu}_{new}$ for $i \in \mathcal{G}_{c_1} \cup \mathcal{G}_{c_2}$. The other groups remain unchanged if their pairwise mean difference is larger than λ_γ. Suppose there is only one pair of such $\mathcal{G}_{c_1}$ and $\mathcal{G}_{c_2}$ that are merged, the total number of of groups is $C-1$ after the merging. Denote the result of 3.11 as L_{C-1} for convenience.

The reason why we merge $\mathcal{G}_{c_1}$ and $\mathcal{G}_{c_2}$ when $|\widehat{\mu}(\mathcal{G}_{c_1}) - \widehat{\mu}(\mathcal{G}_{c_2})| \le \lambda_\gamma$ is clear if we compare L_C and L_{C-1}. The difference between L_C and L_{C-1} is

$$\begin{aligned} L_C - L_{C-1} &= \sum_{c=c_1,c_2}\sum_{i\in\mathcal{G}_c}\sum_{m=1}^{M}\omega_i^{(m,t-1)}\{\gamma_i^{(m,t-1)} - \widehat{\mu}(\mathcal{G}_c)\}^2 \\ &\quad - \sum_{i\in\mathcal{G}_{c_1}\cup\mathcal{G}_{c_2}}\sum_{m=1}^{M}\omega_i^{(m,t-1)}(\gamma_i^{(m,t-1)} - \widehat{\mu}_{new})^2 + \frac{|\mathcal{G}_{c_1}||\mathcal{G}_{c_2}|}{|\mathcal{G}_{c_1}| + |\mathcal{G}_{c_2}|}\lambda_\gamma^2 \\ &= \frac{|\mathcal{G}_{c_1}||\mathcal{G}_{c_2}|}{|\mathcal{G}_{c_1}| + |\mathcal{G}_{c_2}|}\left[\lambda_\gamma^2 - \{\widehat{\mu}(\mathcal{G}_{c_1}) - \widehat{\mu}(\mathcal{G}_{c_2})\}^2\right]. \end{aligned}$$

Clearly, if $|\widehat{\mu}(\mathcal{G}_{c_1}) - \widehat{\mu}(\mathcal{G}_{c_2})| \le \lambda_\gamma$, L_{C-1} is smaller than L_c and the Q_2 is maximized if merging $\mathcal{G}_{c_1}$ and $\mathcal{G}_{c_2}$.

After obtaining $\widehat{\boldsymbol{\mu}}^{(t)}$, the standard deviation $\widehat{\sigma}^{(t)}$ is updated as the square root of

$$L(\widehat{\boldsymbol{\mu}}^{(t)}; \lambda_\gamma) = \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M}\omega_i^{(m,t-1)}(\gamma_i^{(m,t-1)} - \widehat{\mu}_i^{(t)})^2 + \frac{1}{n}\sum_{i_1<i_2}\frac{I(\widehat{\mu}_{i_1}^{(t)} \ne \widehat{\mu}_{i_2}^{(t)})\lambda_\gamma^2}{|\mathcal{G}(i_1)| + |\mathcal{G}(i_2)|}.$$

3.3.4 Initialization and stopping

The model is initialized by assuming $\sigma = 0$ and each γ_i as a cluster with 0 variance by itself, which is equivalent to estimating γ_i as a fixed effect. For computation convenience, we successively update $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ until convergence and obtain $\widehat{\boldsymbol{\beta}}^{(0)}$ and $\widehat{\gamma}_i^{(0)}$. Then the initial value of $\boldsymbol{\beta}$ is set to be $\widehat{\boldsymbol{\beta}}^{(0)}$ and the initial value of μ_i, denoted as $\widehat{\mu}_i^{(0)}$, is set to be equal to $\widehat{\gamma}_i^{(0)}$ for $i = 1, \ldots, n$. If all patients in the transplant center i survive, γ_i may diverge to negative infinity. In this case, we simply set $\widehat{\mu}_i^{(0)}$ to be $\min\{\widehat{\mu}_i^{(0)} | \sum_{k=1}^{N_i} Y_{ik} \neq -N_i, i = 1, \ldots, n\}$. Similarly, if all patients in the transplant center i are dead, we set $\widehat{\mu}_i^{(0)}$ to be $\max\{\widehat{\mu}_i^{(0)} | \sum_{k=1}^{N_i} Y_{ik} \neq N_i, i = 1, \ldots, n\}$.

Different initial values may result into very different clustering trees if the random effects from different groups are merged together incorrectly in the early stage. Our suggestion is adding some random disturbance on $\widehat{\mu}_i^{(0)}$, repeating the clustering procedure multiple times and choosing the best tree with largest information criterion. The random disturbance is generated from normal distribution $N\{0, \widehat{V}(\widehat{\gamma}_i^{(0)})\}$ where $\widehat{V}(\widehat{\gamma}_i^{(0)})$ is the estimated variance of $\widehat{\gamma}_i^{(0)}$. Our experience shows the performance and stability of the algorithm is significantly improved by repeating the procedure multiple times.

Following Booth and Hobert (1999), we stop the EM algorithm at iteration t if

$$\max_l \frac{|\theta_l^{(t)} - \theta_l^{(t-1)}|}{|\theta_l^{(t-1)}| + 0.001} < 0.001,$$

where θ_l is the lth entry in $\boldsymbol{\theta}$.

3.4 Tuning Parameters

3.4.1 Selecting λ_β

There are two tuning parameters in our model, λ_β and λ_γ. Denote $\widehat{\boldsymbol{\theta}}(\lambda_\beta, \lambda_\gamma)$ as the estimation with tuning parameter λ_β and λ_γ and denote $\widehat{\boldsymbol{\mu}}(\lambda_\beta, \lambda_\gamma)$, $\widehat{\sigma}^2(\lambda_\beta, \lambda_\gamma)$ and $\widehat{\boldsymbol{\beta}}(\lambda_\beta, \lambda_\gamma)$ as the corresponding part of $\widehat{\boldsymbol{\theta}}(\lambda_\beta, \lambda_\gamma)$. Then $\widehat{\boldsymbol{\theta}}(\lambda_\beta, \infty)$ is the estimation when $\lambda_\gamma = \infty$. The model with $\lambda_\gamma = \infty$ is equivalent to a regular GLMM assuming the random effects to follow the Gaussian distribution.

We propose to select the λ_β first and select λ_γ next. The λ_β is selected by fitting a regular GLMM on a grid of values and the best λ_β is selected by BIC.

$$\text{BIC}(\lambda_\beta) \quad = \quad -2l_{mar}\{\widehat{\boldsymbol{\theta}}(\lambda_\beta, \infty)\} + \|\widehat{\boldsymbol{\beta}}(\lambda_\beta, \infty)\|_0 \log N \tag{3.12}$$

where $\|\widehat{\boldsymbol{\beta}}(\lambda_\beta, \infty)\|_0$ is the number of nonzero elements in $\widehat{\boldsymbol{\beta}}(\lambda_\beta, \infty)$. The log marginal likelihood is

$$l_{mar}\{\widehat{\boldsymbol{\theta}}(\lambda_\beta, \infty)\} = \sum_{i=1}^{n} \log \int \prod_{k=1}^{N_i} f\{Y_{ik}|\boldsymbol{X}_{ik}, \gamma_i; \widehat{\boldsymbol{\beta}}(\lambda_\beta, \infty)\} f\{\gamma|\widehat{\mu}_i(\lambda_\beta, \infty), \widehat{\sigma}(\lambda_\beta, \infty)\} d\gamma_i$$

and can be well approximated by

$$\sum_{i=1}^{n} \log \sum_{m=1}^{M} \frac{1}{\sqrt{\pi}} v_m \prod_{k=1}^{N_i} f\{\boldsymbol{Y}_{ik}|\boldsymbol{X}_{ik}, \gamma_i^{(m)}(\lambda_\beta, \infty); \widehat{\boldsymbol{\beta}}(\lambda_\beta, \infty)\}$$

where $\gamma_i^{(m)}(\lambda_\beta, \infty) = \widehat{\mu}_i(\lambda_\beta, \infty) + \sqrt{2}\widehat{\sigma}(\lambda_\beta, \infty) d_m$. Then we fixed $\widehat{\boldsymbol{\beta}}(\lambda_\beta, \infty)$ and choose λ_γ, or the number of clusters by rules such as Gap statistic or stability selection.

3.4.2 Selecting the number of clusters

3.4.2.1 Gap Statistic

The Gap statistic proposed by Tibshirani et al. (2001) was trying to find the largest gap between $\log(W_C)$ and its expected value under reference distribution, where W_C is the within cluster sum of squares when there are C clusters for $1 \leq C \leq C_{max}$. The maximum number of clusters C_{max} is a constant that we believe it is larger than the true number of components. If $\boldsymbol{\gamma}$ is observed the within cluster sum of squares can be written explicitly as

$$W_C = \sum_{c=1}^{C} \sum_{i \in \mathcal{G}_c} \{\gamma_i - \widehat{\mu}(\mathcal{G}_c)\}^2 = n\widehat{\sigma}_C^2$$

where $\widehat{\mu}(\mathcal{G}_c)$ for $c = 1, \ldots C$ are the estimated cluster means and $\widehat{\sigma}_C^2$ is the estimated variance under the model with C subgroups. The expectation $E\{\log(W_C)\}$ and standard deviation of $\log(W_C)$, denoted as sd_C, are estimated from B reference data sets draw from a reference distribution. We run the same clustering algorithm on each reference data set to obtain multiple replications $\log(W_{C,b}^*)$. The Gap statistic is then defined as

$$Gap(C) = \frac{1}{B} \sum_{b=1}^{B} \log(W_{C,b}^*) - \log(W_C)$$

and C clusters is preferred than $C+1$ if

$$Gap(C) \geq Gap(C+1) - sd_{C+1}\sqrt{1+1/B}.$$

The reference distribution is the uniform distribution with the same range as the real data.

In the case when $\boldsymbol{\gamma}$ are unobserved, the definition of $\log(W_C)$ is not immediate. Possible alternatives are $\log(n\widehat{\sigma}^2_C)$ and the log likelihood $l_{mar}(\widehat{\boldsymbol{\theta}}_C)$ where $\widehat{\boldsymbol{\theta}}_C$ are the estimated parameters under the model with C subgroups and $\widehat{\sigma}^2_C$ is the estimate of σ^2. Correspondingly the expectations should be replaced by the means of $\log(n\widehat{\sigma}^{2*}_{C,b})$ and log marginal likelihood $l_{mar}(\widehat{\boldsymbol{\theta}}^*_{C,b})$ for $b = 1, \ldots, B$. Since $\boldsymbol{\gamma}$ is unobserved its range is estimated by the minimum and maximum of the initial values $\widehat{\gamma}^{(0)}_i$ $(i = 1, \ldots, n)$. Recall that the initial value of $\boldsymbol{\gamma}$ are estimated by treating them as fixed effects, so we expect the range is large enough to approximately cover the true range.

We try 3 different implementations of the Gap statistics, Gap_1, Gap_2, and Gap_3. The Gap_1 and Gap_2 are based on $\log(n\widehat{\sigma}^{2*}_C)$. The difference is the way to estimate the expectations. After generating the reference data $\gamma^*_{i,b}$ $(i = 1, \ldots, n; b = 1, \ldots, B)$, the Gap_1 runs the clustering algorithm directly on $\boldsymbol{\gamma}^*_b$ and calculates the mean of $\log(n\widehat{\sigma}^{2*}_{C,b})$. The Gap_2, however, generates $\boldsymbol{Y}^*_{i,b}$ based on $\boldsymbol{X}_i$, $\widehat{\boldsymbol{\beta}}(\lambda_\beta)$ and $\gamma^*_{i,b}$ for $i = 1, \ldots, n$ and $b = 1, \ldots, B$. Then it fits the GLMM on each of the generated dataset with $\widehat{\boldsymbol{\beta}}(\lambda_\beta)$ and $\gamma^*_{i,b}$ fixed and calculates the mean of $\log(n\widehat{\sigma}^{2*}_{C,b})$ where the $\widehat{\sigma}^{2*}_{C,b}$ is estimated from the GLMM and is different from that used in Gap_1. The Gap_3 is based on $l_{mar}(\widehat{\boldsymbol{\theta}}_C)$ and the expectation is estimated by the mean of $l_{mar}(\widehat{\boldsymbol{\theta}}^*_{C,b})$ estimated from GLMM for $b = 1, \ldots, B$. In our simulation, the number of reference datasets B for Gap_1 is 1000 and that for Gap_2 and Gap_3 is chosen to be 200. To reduce computation time, the number of Gauss-Hermite quadrature points are chosen to be 20 and the the fixed effects coefficients are fixed at $\widehat{\boldsymbol{\beta}}(\lambda_\beta)$ for model fitting on the reference datasets.

3.4.2.2 Stability Selection

Stability selection (Meinshausen and Bühlmann, 2010) is originally proposed for variable selection. Wang (2010) extends it to select the number of clusters by a modified cross

validation with the goal of minimizing the clustering instability. Fang and Wang (2012) use the same idea but they evaluate the clustering instability by resampling rather than cross validation. Denote ψ_1 and ψ_2 as two clustering rules, which map a data point to its cluster label. The clustering distance is defined as

$$d(\psi_1, \psi_2) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} [|I\{\psi_1(\boldsymbol{Y}_i) = \psi_1(\boldsymbol{Y}_j)\} - I\{\psi_2(\boldsymbol{Y}_i) = \psi_2(\boldsymbol{Y}_j)\}|]$$

where $I\{\psi_1(\boldsymbol{Y}_i) = \psi_1(\boldsymbol{Y}_j)\}$ is 1 if $\psi_1(\boldsymbol{Y}_i) = \psi_1(\boldsymbol{Y}_j)$. Otherwise, $I\{\psi_1(\boldsymbol{Y}_i) = \psi_1(\boldsymbol{Y}_j)\}$ equals to 0. Generate B pairs of independent bootstrap data $\{(\boldsymbol{Y}^*_{b,1}, \boldsymbol{X}^*_{b,1}), (\boldsymbol{Y}^*_{b,2}, \boldsymbol{X}^*_{b,2})\}$ for $b = 1, \ldots, B$. Run the clustering algorithm on both of the datasets for each pair. We obtain clustering rules $\psi^*_{b,1,C}$ from $(\boldsymbol{Y}^*_{b,1}, \boldsymbol{X}^*_{b,1})$ and $\psi^*_{b,2,C}$ from $(\boldsymbol{Y}^*_{b,2}, \boldsymbol{X}^*_{b,2})$ for $C = 2, \ldots, C_{max}$. Then the clustering instability is defined as

$$s_B(C) = \frac{1}{B} \sum_{b=1}^{B} d(\psi^*_{b,1,C}, \psi^*_{b,2,C}).$$

The clustering distance is evaluated on the original data set instead of the bootstrap samples. The number of clusters is chosen to be

$$\widehat{C} = \arg \min_{2 \le C \le C_{max}} s_B(C).$$

Since the clustering instability is always 0 when $C = 1$, stability selection can only choose C for $C \geq 2$. The number of bootstrap sample B we use here is 200 and the number of Gauss-Hermite quadrature points are chosen to be 20. The fixed effects coefficients are fixed at $\widehat{\boldsymbol{\beta}}(\lambda_\beta)$ for model fitting on the bootstrapping datasets, in the same way as the Gap statistic.

3.4.2.3 Information Criteria

Ma and Huang (2017) uses the mBIC to select the number of clusters. If mBIC works, it is possible select the fixed effects and random effects jointly as Hui et al. (2016). Combining the BIC together with mBIC, we obtain the following joint rule

$$\begin{aligned} m\text{BIC}_d(\lambda_\beta, \lambda_\gamma) &= -2l_{mar}\{\widehat{\boldsymbol{\theta}}(\lambda_\beta, \lambda_\gamma)\} + \|\widehat{\boldsymbol{\beta}}(\lambda_\beta, \lambda_\gamma)\|_0 \log N \\ &\quad + d\ C\{\widehat{\boldsymbol{\mu}}(\lambda_\beta, \lambda_\gamma)\}\ \log(n) \log\{\log(n)\} \end{aligned}$$

where $\|\widehat{\boldsymbol{\beta}}(\lambda_\beta, \lambda_\gamma)\|_0$ is the number of nonzero elements in $\widehat{\boldsymbol{\beta}}(\lambda_\beta, \lambda_\gamma)$ and $C\{\widehat{\boldsymbol{\mu}}(\lambda_\beta, \lambda_\gamma)\}$ is the number of unqiue values of $\widehat{\boldsymbol{\mu}}(\lambda_\beta, \lambda_\gamma)$. The d in $m\text{BIC}_d$ is a constant chosen to be 5 or 10 by Ma and Huang (2017).

AIC_{mix} is original proposed to select the order in a normal mixture model. Our model estimation $\widehat{\boldsymbol{\theta}}(\lambda_\beta, \lambda_\gamma)$ can also be viewed as a normal mixture model for comparison purposes. Let $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\mu}}$ and $\widehat{\sigma}$ be the corresponding parts of $\widehat{\boldsymbol{\theta}}(\lambda_\beta, \lambda_\gamma)$. We suppress their dependency on tuning parameter for easy of notation. Denote $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_C)^T$ as the C unique values of $\widehat{\boldsymbol{\mu}}$ and use $\tilde{\mu}$ as the components means of a normal mixture. Denote $\tilde{\pi}_c = |\mathcal{G}|_c/n$ and $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_C)^T$ as component weights and the common component standard deviation is set to be $\widehat{\sigma}(\lambda_\beta, \lambda_\gamma)$. Then the AIC_{mix} under our framework is defined as

$$-2\sum_{i=1}^{n}\log\left\{\sum_{c=1}^{C}\tilde{\pi}_c f(\boldsymbol{Y}_i|\boldsymbol{X}_i;\widehat{\boldsymbol{\beta}},\tilde{\mu}_c,\widehat{\sigma})\right\} \tag{3.13}$$
$$+2tr\left\{\boldsymbol{I}_{comp}(\widehat{\boldsymbol{\tau}};\widehat{\boldsymbol{\beta}},\tilde{\boldsymbol{\mu}},\widehat{\sigma}^2)\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}},\tilde{\boldsymbol{\mu}},\widehat{\sigma}^2)\right\} + 2\sum_{c=1}^{C}\frac{\sum_{i=1}^{n}\widehat{\tau}_{ic}}{\widehat{\tau}_{ic}^2}$$

where

$$f(\boldsymbol{Y}_i|\boldsymbol{X}_i;\widehat{\boldsymbol{\beta}},\tilde{\mu}_c,\widehat{\sigma}) = \int f(\boldsymbol{Y}_i|\boldsymbol{X}_i,\gamma;\widehat{\boldsymbol{\beta}})f(\gamma|\tilde{\mu}_c,\widehat{\sigma})d\gamma$$

is the conditional likelihood $\boldsymbol{Y}_i$ given $\boldsymbol{X}_i$. The matrix $\widehat{\boldsymbol{\tau}}$ has $\widehat{\tau}_{ic}$ as its ith row and cth column element $(i = 1, \ldots, n; c = 1, \ldots, C)$, denoting the posterior probability for observation $(\boldsymbol{Y}_i, \boldsymbol{x}_i^T)^T$ belongs to component c. The estimation of τ_{ic} is

$$\widehat{\tau}_{ic} = \frac{\tilde{\pi}_c f(\boldsymbol{Y}_i|\boldsymbol{X}_i;\widehat{\boldsymbol{\beta}},\tilde{\mu}_c,\widehat{\sigma})}{\sum_{c=1}^{C}\tilde{\pi}_c f(\boldsymbol{Y}_i|\boldsymbol{X}_i;\widehat{\boldsymbol{\beta}},\tilde{\mu}_c,\widehat{\sigma})}.$$

The matrix $\boldsymbol{I}(\widehat{\boldsymbol{\beta}},\tilde{\boldsymbol{\mu}},\widehat{\sigma}^2)$ is the observed Fisher information, defined as the second derivative of

$$-\sum_{i=1}^{n}\log\left\{\sum_{c=1}^{C}\widehat{\pi}_c f(\boldsymbol{Y}_i|\boldsymbol{X}_i;\widehat{\boldsymbol{\beta}},\tilde{\mu}_c,\widehat{\sigma}_c)\right\}$$

while $\boldsymbol{I}_{comp}(\widehat{\boldsymbol{\tau}};\widehat{\boldsymbol{\beta}},\tilde{\boldsymbol{\mu}},\widehat{\boldsymbol{\sigma}}^2)$ is defined as the second derivative of

$$-\sum_{i=1}^{n}\sum_{c=1}^{C}\widehat{\tau}_{ic}\left\{\log(\tilde{\pi}_c) + \log f(\boldsymbol{Y}_i|\boldsymbol{X}_i;\widehat{\boldsymbol{\beta}},\tilde{\mu}_c,\widehat{\sigma}_c)\right\}.$$

When $\boldsymbol{\beta}$ is fixed at $\widehat{\boldsymbol{\beta}}(\lambda_\beta, \infty)$ we also try the following naive BIC to select the number of subgroups for comparison purpose

$$\sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \tilde{\pi}_c f(\boldsymbol{Y}_i | \boldsymbol{X}_i; \widehat{\boldsymbol{\beta}}(\lambda_\beta, \infty), \tilde{\mu}_c, \widehat{\sigma}) \right\} + (3C-1)\log(n).$$

3.4.2.4 Cross-validation

The cross-validation (Stone, 1974) is a popular and powerful approach for model selection. In order to predict on new observations in the clustering context, we need to define a normal mixture distribution in the same way as the AIC_{mix}. Randomly divide the the n transplant centers into 5 folds. Use the bth fold as test set and the remaining 4 folds as training set for $b = 1, \ldots, 5$, we obtain 5 pairs of datasets $\{\mathcal{A}_{train,b}, \mathcal{A}_{test,b}\}_{b=1}^{5}$.

Fit our clustering algorithm on $\mathcal{A}_{train,b}$ with the patient level coefficients fixed at $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\lambda_\beta, \infty)$, we obtain $\tilde{\pi}_c(\mathcal{A}_{train,b})$, $\tilde{\mu}_c(\mathcal{A}_{train,b})$ and $\widehat{\sigma}(\mathcal{A}_{train,b})$ in the similar way as in AIC_{mix} for $c = 1, \ldots, C$. The log marginal likelihood on test set is

$$l_{C,b,mar} = \sum_{i \in \mathcal{A}_{test,b}} \log \int \prod_{k=1}^{N_i} f(Y_{ik} | \boldsymbol{X}_{ik}, \gamma_i; \widehat{\boldsymbol{\beta}}) \\ \sum_{c=1}^{C} \tilde{\pi}_c(\mathcal{A}_{train,b}) f\{\gamma | \tilde{\mu}_c(\mathcal{A}_{train,b}), \widehat{\sigma}(\mathcal{A}_{train,b})\} d\gamma_i.$$

The number of clusters is chosen as

$$\widehat{C} = \arg \max_{1 \le C \le C_{max}} \sum_{b=1}^{5} l_{C,b,mar}.$$

3.4.2.5 Mean integrated squared error

We view the clustering problem as a density estimation problem and we select the number of clusters by minimizing the Mean Integrated Squared Error (MISE). Assume the current density function is a C components normal mixture

$$\widehat{f}(\gamma \mid \widehat{\boldsymbol{\theta}}_C) = \sum_{c=1}^{C} \frac{|\mathcal{G}_c|}{n} f(\gamma \mid \tilde{\mu}_c, \widehat{\sigma})$$

when there are C clusters, where $\tilde{\mu}_c$ is the mean of the cth cluster, $\widehat{\sigma}$ is the common standard deviation for all clusters and f is a normal density function. Denoting the true distribution as $g(\gamma)$, the MISE is defined as

$$\int \{g(\gamma) - \widehat{f}(\gamma \mid \widehat{\boldsymbol{\theta}}_C)\}^2 d\gamma = \int g(\gamma)^2 d\gamma - 2 \int g(\gamma) \widehat{f}(\gamma \mid \widehat{\boldsymbol{\theta}}_C) d\gamma + \int \widehat{f}(\gamma \mid \widehat{\boldsymbol{\theta}}_C)^2 d\gamma.$$

The third term can be written explicitly as

$$\sum_{c_1=1}^{C}\sum_{c_2=1}^{C}\frac{|\mathcal{G}_{c_1}||\mathcal{G}_{c_2}|}{n^2}\frac{1}{\sqrt{4\pi\widehat{\sigma}_C^2}}\exp\left\{-\frac{(\widehat{\mu}_{c_1}-\widehat{\mu}_{c_2})^2}{4\widehat{\sigma}_C^2}\right\}.$$

The second term can be estimated by

$$-\frac{2}{n}\sum_{i=1}^{n}\widehat{f}(\widehat{\gamma}_i^{(0)}\mid\widehat{\boldsymbol{\theta}}_C)$$

where $\widehat{\gamma}_i^{(0)}$ is the initial value for γ_i and is estimated by treating γ_i as fixed effects.

3.5 Simulation Studies

Denote the overall mean of $\widehat{\gamma}_i$ of real data as μ_R, and μ_R is approximately -1.26. We randomly generate γ_i $(i=1,\ldots,300)$ from the following normal mixture models

Model 0:	$N(\mu_R, 1^2)$,
Model 1:	$0.5\ N(\mu_R-1,1^2)+0.5\ N(\mu_R+1,1^2)$,
Model 2:	$0.5\ N(\mu_R-1.5,1^2)+0.5\ N(\mu_R+1.5,1^2)$,
Model 3:	$0.5\ N(\mu_R-2,1^2)+0.5\ N(\mu_R+2,1^2)$,
Model 4:	$1/3\ N(\mu_R-1,0.5^2)+1/3\ N(\mu_R,0.5^2)+1/3\ N(\mu_R+1,0.5^2)$,
Model 5:	$1/3\ N(\mu_R-1.5,0.5^2)+1/3\ N(\mu_R,0.5^2)+1/3\ N(\mu_R+1.5,0.5^2)$,
Model 6:	$1/3\ N(\mu_R-2,0.5^2)+1/3\ N(\mu_R,0.5^2)+1/3\ N(\mu_R+2,0.5^2)$.

Then we generate the number of patients per transplant center as the rounded sum of two random numbers from $Poisson(5)$ and $Exponential(45)$ to mimic the skewed distribution of the N_i in the real data. The response Y_{ik} is a binary variable generated using (3.4) with $P(Y_{ik}=1)=\{1+\exp(-\xi_{ik})\}^{-1}$, where $\xi_{ik}=\boldsymbol{X}_{ik}^{\mathrm{T}}\boldsymbol{\beta}+\gamma_i$. The patient level covariates $\boldsymbol{X}$ is a matrix with $\sum_{i=1}^{n}N_i$ rows and ten columns. Each column of $\boldsymbol{X}$ is generated independently from a normal distribution with zero means. The standard deviation of column j is $\exp(j)/\{\sum_{j=1}^{4}\exp(j)+6\}$ for $1\le j\le 4$ and the standard deviations of the following six columns are all equal to $1/\{\sum_{j=1}^{4}\exp(j)+6\}$. The fixed effects $\boldsymbol{\beta}$ is a vector of length 10. Its first four elements are 2, corresponding to the important effects, while the following six elements are 0, corresponding to the redundant effects. The candidate set of λ_β is

$(0.0, 0.00125, 0.0025, 0.005, 0.01, 0.02, 0.04, 0.08, 0.16)$ and the simulations show these values can approximately result in models with 10 to 2 nonzero elements in $\widehat{\boldsymbol{\beta}}$. For each λ_β, we try 10 random initial values.

200 independent datasets are generated for each of the model 0–6, and the clustering procedure is run 90 times in total for each dataset. Some dendrograms of model 1–6 are shown in Fig. 3.1. The model selection results for $\boldsymbol{\beta}$ are shown in Table 3.2. As we can see, the percentage of correct selection is very high. This is understandable considering the number of total patients N is 269386. The high proportion of correct selection and the relatively precise estimation of $\boldsymbol{\beta}$ is the reason we select $\boldsymbol{\beta}$ before selecting the number of clusters.

The cluster selection results are shown in Table 3.3, 3.4 and 3.5. A summary of all the methods is given below.

- The $m\text{BIC}_d$ over-select in all cases for all choices of d.
- The AIC_{mix} and naiveBIC always prefer one cluster model. These two rules are defined in the mixture model case. They should work better if $\tilde{\boldsymbol{\theta}}$ are estimated by maximizing the likelihood of a mixture model. However, we are maximizing a totally different model and the log marginal likelihood is not "properly" maximized and thus, these two rules always select the one cluster model.
- Stability selection can select the true number of clusters when the clusters are well separated such as model 3 and 6.
- Gap_1 and Gap_2 only work for model 1–3 when there are only 2 clusters.
- Gap_3 does not work at all.
- The MISE and cross-validation work better than the other methods.

Although MISE and cross-validation are the best among all the available choices, their performance is not perfect. We run some additional simulations to evaluate the clustering results by the Rand Index (Rand, 1971) when the true number of clusters is given. The

Rand Index is defined as

$$RI = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \left[\left| I\{\widehat{\psi}(\boldsymbol{Y}_i) = \widehat{\psi}(\boldsymbol{Y}_j)\} - I\{\psi_0(\boldsymbol{Y}_i) = \psi_0(\boldsymbol{Y}_j)\} \right| \right]$$

where $\widehat{\psi}$ is the estimated clustering rule and ψ_0 is the true clustering rule. Table 3.1 shows the clustering method can approximately recover the true clusters but we do not have a good method to select the right λ_γ. Developing a good rule for selecting the number of clusters is still an open question.

Table 3.1 The minimum, 1st quartile, median, mean, 3rd quartile and maximum of the Rand Index for Model 1–6 based on 200 simulations when true number of clusters is given.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Model 1	0.50	0.81	0.88	0.82	0.90	0.95
Model 2	0.50	0.92	0.94	0.93	0.95	0.99
Model 3	0.89	0.97	0.98	0.97	0.99	1.00
Model 4	0.61	0.75	0.78	0.78	0.81	0.89
Model 5	0.71	0.85	0.87	0.87	0.89	0.95
Model 6	0.87	0.93	0.94	0.94	0.95	0.98

Table 3.2 Summary of the fixed effects selected by BIC in model 0–6 based on 200 simulations. Correct selection: frequency of the correct selecting all important effects and set all redundant effects to 0; Miss nonzero covaraties: frequency of at least one important effects is not selected; Average model size: the average number of variables selected.

	Correct selection	Miss nonzero covariates	Average model size
Modle 0	198	0	4.01
Modle 1	198	0	4.01
Modle 2	200	0	4
Modle 3	200	0	4
Modle 4	199	0	4.005
Modle 5	200	0	4
Modle 6	200	0	4

Table 3.3 Number of clusters selected by each of the methods in model 0. The true number of clusters is 1.

		1	2	3	4	5+
Model 0	mBIC_1	0	0	0	0	200
	mBIC_5	0	0	0	1	199
	mBIC_10	0	0	42	74	84
	naive BIC	200	0	0	0	0
	AICmix	200	0	0	0	0
	Stability	0	32	4	6	158
	Gap1	200	0	0	0	0
	Gap2	200	0	0	0	0
	Gap3	200	0	0	0	0
	CV	62	46	36	26	30
	MISE	166	25	5	3	1

3.6 Data Analysis

3.6.1 Background

Renal failure is one of the most common and severe diseases in the United States. In 2013, a total of 117,162 new cases were reported (www.USRDS.org). Kidney transplantation, a primary therapy for end stage renal disease, is a complicated procedure typically involving transplant surgeons and physicians, coordinators, social workers, financial counselors, nutritionists, psychologists and referring physicians. The quality of care delivered by a transplant center is often assessed by patient survival, such as the 5 year post-transplant survival rate.

Our motivating data are obtained from the Organ Procurement and Transplantation Network, administered by the U.S. Department of Health and Human Services. The data system includes data on all donors, wait-listed candidates, and transplant recipients in the U.S. Included in the analysis are adult renal failure patients ($\geq$ 18 years of age) who underwent deceased donor kidney transplantation between January 1987 and December 2008. This cohort includes $N = 269,386$ patients receiving kidney transplants from a total of $n = 296$ centers. The number of transplants performed by a center, N_i, has a highly skewed distribution. Most centers performed a few hundred cases of kidney transplantation,

but there are centers that took over 5000 cases. The patient level response is the 5-year survival status (1=death and -1=survival) and there is no censoring due to routine and rigorous tracking of the patients. The overall 5-year failure rate is 27.59%.

An important patient level covariate that is directly related to the success of kidney transplants is x_1 = cold ischemic time, which is the time that the donor kidney was kept in a refrigerator before being received by the patient. Other patient level covariates include x_2 = age at transplantation and x_3 = sex of the patient (1 = male, 0 = female), while x_4–x_6 are indicators for BMI in the intervals (22, 25], (25, 30] and 30+ respectively. Since the data were collected over a time span of two decades, it is possible that the technology used in transplant surgeries has improved over time, which also affects the patient level outcome. Therefore, in addition to the other covariates described above, we also include time effects into the model. Using cases before 1990 as the baseline, covariates x_7–x_{10} are indicators for cases performed in 1990–1994, 1995–1999, 2000–2003 and 2004–2008 respectively. See Table 3.6 for a summary of the categorical variables including sex of the patient, BMI and the surgery time.

3.6.2 Model fitting

We run the model on the real data for different choices of λ_β while letting the transplant center effects follow a homogeneous Gaussian distribution. Then we select the the model with the largest BIC defined as in (3.12). The BIC selects nine important variables except for x_3, sex of the patient. The summary of the patient level effects are given at Table 3.7. The variable x_3 is also the least significant variable in our previous work. To confirm the legitimate of selecting λ_β before λ_γ, we plot the patient level effects along with the λ_γ in Fig. 3.2. From this plot, some jump is observed at the early stage but $\widehat{\boldsymbol{\beta}}$ becomes very stable against λ_γ after that. This justifies that we can select the variables in $\boldsymbol{\beta}$ first and select the number of clusters next.

Then we choose the number of clusters using cross validation. The five-fold cross validation chooses 2 clusters. The means of the two clusters are -3.04 and -0.97 and the number of patients for the two clusters are 5 and 290 respectively. The common standard deviation

is 0.25. The overall estimation is very close to our estimation in previous work. Apply the Gap_1, Gap_2 and Gap_3 to the real data and only one cluster is chosen. This is not surprising given their performance on the simulations. Apply the MISE to the real data and still only one cluster is chosen. The MISE and the Gap statistic have the tendency to under select when there are multiple clusters but they are not easily distinguishable. Combining with our research from previous work, we prefer to select 2 clusters. Another strong evidence suggesting the existence of two clusters in the data is that all the 5 training models in the five fold cross validation give similar estimation when the number of clusters are two.

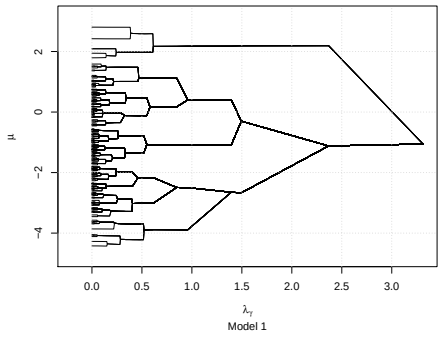
We fit the clustering algorithm with the nine selected variables on the real data and obtain the dendrogram as shown in Fig. 3.3. The dendrogram displays how $\widehat{\mu}_i$ (i=1,. . . ,n) change along with λ_γ. A smaller μ stands for outperforming and a larger μ stands for underperforming. From the dendrogram, we can see a small subgroup with 5 transplant centers stands out and does not merge with the others until only 1 big cluster remains. We may call it the outperforming group. The outperforming group corresponds to 5 transplant centers, #10, #290, #287, #264 and #265. Table 3.8 shows the sample size and death rate of the 5 transplant centers. The center #264 and center #265 with 0 death rate are set to be equal to the $\widehat{\mu}_{287}$ at initialization stage. Compared to our previous work, the center #28 was previously detected as outperforming but it does not belong to the current outperforming group. The center #264, #265 and #290 were not detected as outperforming in our previous work. Our previous False Discovery Rate control approach is more conservative and it only selects a transplant center when the sample size is large and the evidence is strong. In contrast, our current approach is more sensitive to death rate.

3.7 Discussion

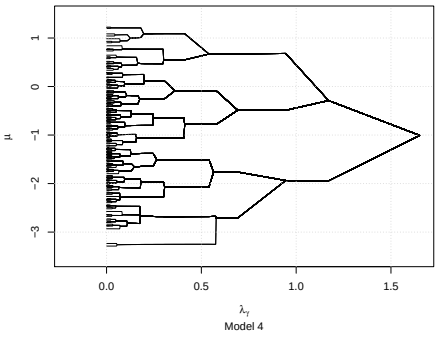
We propose a method that performs subgroup analysis on random effects of a GLMM while it can also select important fixed effects. We demonstrate that our algorithm can successfully detect the subgroups of the random effects. It can also give the clustering dendrogram that shows the structure of the heterogeneity. We use l_0 penalty which is different but motivated from the previous research in Wu et al. (2016) and Ma and Huang

(2017). Our algorithm gives similar clustering tree to the algorithms of Wu et al. (2016) and Ma and Huang (2017) but requires much less computation. We compare several methods of selecting the number of clusters. The Mean Integrated Squared Error and cross validation perform better than other methods such as the Gap statistic and stability selection.

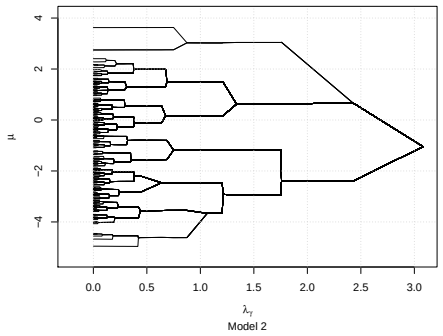
We apply our method to the kidney transplant center evaluation data. The variable selection procedure suggests sex of the patients is not an important predictor for the patient-level surviving outcome. While the other predictors, such as cold ischemic time, age at transplantation, BMI and year of transplantation are important predictors. Two clusters are discovered among the transplant centers. A small cluster with 5 transplant centers stands out in the dendrogram and they outperform the other 290 transplant centers. The 5 transplant centers treated 269 patients in total and 11 of them die within 5 years after the surgery. The selected 5 transplant centers are different from the ones detected in our previous work. Our current method is more sensitive to death rate and the 5 selected transplant centers have the lowest death rate.



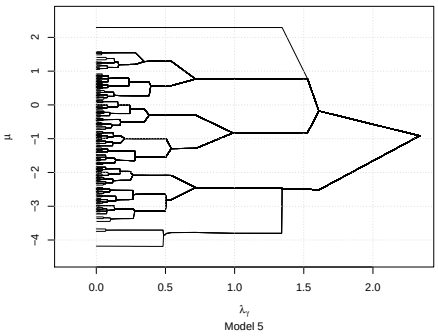
(a) Model 1



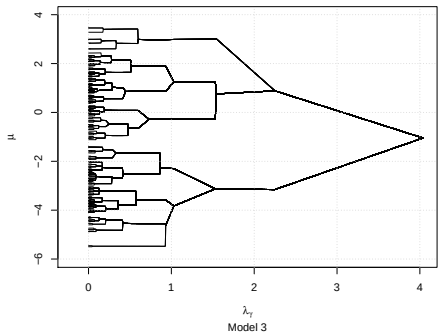
(d) Model 4



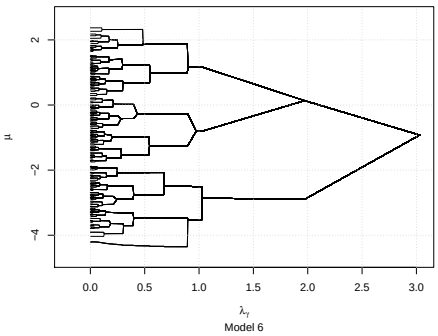
(b) Model 2



(e) Model 5



(c) Model 3



(f) Model 6

Figure 3.1 The dendrograms of Model 1–6. The x-axis is the λ_γ and the y-axis is $\widehat{\mu}_i$. Each of the lines represents how $\widehat{\mu}_i$ of transplant center i is changing along λ_γ.

Table 3.4 Number of clusters selected by each of the methods in model 1, 2 and 3 based on 200 simulations. The true number of clusters is 2.

		1	2	3	4	5+
Model 1	mBIC_1	0	0	0	0	200
	mBIC_5	0	0	0	0	200
	mBIC_10	0	1	26	72	101
	naive BIC	200	0	0	0	0
	AICmix	200	0	0	0	0
	Stability	0	14	2	14	170
	Gap1	200	0	0	0	0
	Gap2	194	6	0	0	0
	Gap3	200	0	0	0	0
	CV	0	59	60	51	30
	MISE	171	12	4	6	7
Model 2	mBIC_1	0	0	0	0	200
	mBIC_5	0	0	0	0	200
	mBIC_10	0	4	19	64	113
	naive BIC	200	0	0	0	0
	AICmix	200	0	0	0	0
	Stability	0	83	48	8	61
	Gap1	83	116	1	0	0
	Gap2	131	69	0	0	0
	Gap3	200	0	0	0	0
	CV	0	96	50	28	26
	MISE	7	108	59	12	14
Model 3	mBIC_1	0	0	0	0	200
	mBIC_5	0	0	0	0	200
	mBIC_10	0	8	17	60	115
	naive BIC	200	0	0	0	0
	AICmix	200	0	0	0	0
	Stability	0	160	38	1	1
	Gap1	0	200	0	0	0
	Gap2	2	198	0	0	0
	Gap3	200	0	0	0	0
	CV	0	103	42	23	32
	MISE	0	129	49	12	10

Table 3.5 Number of clusters selected by each of the methods in model 4,5 and 6 based on 200 simulations. The true number of clusters is 3.

		1	2	3	4	5+
Model 4	mBIC_1	0	0	0	0	200
	mBIC_5	0	0	0	0	200
	mBIC_10	0	1	57	50	92
	naive BIC	200	0	0	0	0
	AICmix	200	0	0	0	0
	Stability	0	20	8	5	167
	Gap1	157	41	2	0	0
	Gap2	200	0	0	0	0
	Gap3	200	0	0	0	0
	CV	11	107	38	24	20
	MISE	132	31	25	9	3
Model 5	mBIC_1	0	0	0	0	200
	mBIC_5	0	0	0	0	200
	mBIC_10	0	0	79	61	60
	naive BIC	200	0	0	0	0
	AICmix	200	0	0	0	0
	Stability	0	36	22	32	110
	Gap1	127	41	27	5	0
	Gap2	200	0	0	0	0
	Gap3	200	0	0	0	0
	CV	0	46	89	48	17
	MISE	32	36	76	42	14
Model 6	mBIC_1	0	0	0	0	200
	mBIC_5	0	0	0	0	200
	mBIC_10	0	0	142	32	26
	naive BIC	200	0	0	0	0
	AICmix	200	0	0	0	0
	Stability	0	8	113	71	8
	Gap1	152	0	47	1	0
	Gap2	197	3	0	0	0
	Gap3	200	0	0	0	0
	CV	0	0	157	36	7
	MISE	0	0	156	30	14

Table 3.6 Summary of the categorical variables sex of the patient, BMI and surgery time.

Variables	Levels	Number of patients
Sex	Male	161791
	Female	107595
BMI	≤ 22	47729
	(22, 25]	51955
	(25, 30]	122739
	30+	46963
Surgery time	1987–1989	19569
	1990–1994	49837
	1995–1999	59421
	2000–2003	57818
	2004–2008	82741

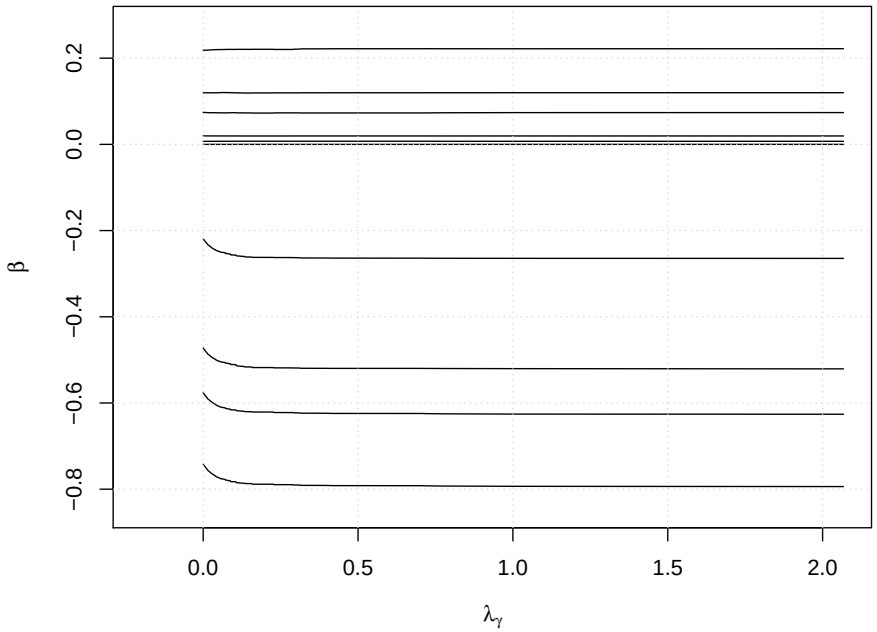


Figure 3.2 The patient level effects $\widehat{\boldsymbol{\beta}}$ (y-axis) of the real data is very stable against the tuning parameter λ_γ (x-axis) except for some fluctations when λ_γ is very small.

Table 3.7 U.S. Organ Procurement and Transplantation Network data analysis: estimated fixed effect coefficients, standard errors, z-values and p-values. .

	Estimate	Std. Error	z-value	p-value
x_1	0.019518	0.0003023	64.565	<1e-99
x_2	0.007142	0.0002111	33.8323	<1e-99
x_3	0.0	NA	NA	NA
x_4	0.074186	0.0153448	4.8346	<1e-5
x_5	0.120468	0.012716	9.4737	<1e-20
x_6	0.223233	0.0145818	15.309	<1e-52
x_7	-0.268844	0.0144038	-18.6648	<1e-77
x_8	-0.525684	0.0125537	-41.8748	<1e-99
x_9	-0.631029	0.0136499	-46.2296	<1e-99
x_{10}	-0.799043	0.0126721	-63.0553	<1e-99

Table 3.8 The outperforming cluster in the kidney transplant data selected by cross validation.

Center ID	Sample.Size	Death.Rate
#10	125	0.056
#264	3	0.000
#265	4	0.000
#287	114	0.026
#290	23	0.043

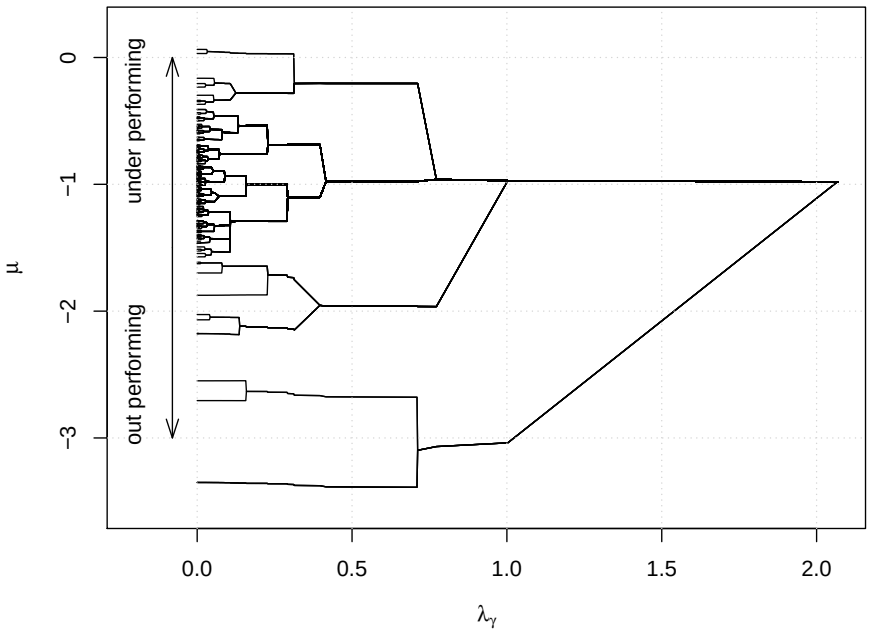


Figure 3.3 The dendrogram of the transplant center effects (y-axis) of the real data against the tuning parameter λ_γ (x-axis). As the increase of λ_γ, the number of clusters decreases from n to 1.

CHAPTER 4. SUMMARY AND FUTURE WORK

4.1 Summary

GLMM typically assumes that the random effects follow a homogeneous Gaussian distribution. When the true random effects are actually drawn from a population consisting of subgroups, the predicted random effects will be deceptively homogeneous, which in turn falsely encourages the usage of homogeneous assumption on random effects. Models without homogeneity assumptions on the random effects are investigated under the framework of GLMM in this dissertation.

The first method assumes the random effects follow a normal mixture distribution. Each component of the normal mixture corresponds to a cluster in the data and each data point has a positive probability of belonging to any of the clusters. This method can consistently estimate the true parameters. A sequential testing approach is developed to determine the number of components in the mixture model. The testing approach is shown to be very powerful in detecting heterogeneity in the random effects.

The second method seeks to find the subgroups in random effects under a different clustering framework. The model automatically groups the random effects with similar means together by penalizing on the pairwise differences. Each random effect is allowed to be an individual cluster when the tuning parameter is 0. All random effects become homogeneous when the tuning parameter is infinity. By choosing a tuning parameter in between 0 and infinity, the random effects are assigned to different groups.

Both methods try to find the clusters in the random effects. The first model has a more solid theoretical foundation. The model consistency is shown and the convergence rate is obtained. The asymptotic distribution of the test statistic is obtained for the sequential testing procedure and the chance of making mistakes is controlled at a prespecified level. The second model investigates the problem under a more flexible framework, does not require a prespecified number of clusters and requires much less computation time.

Both methods are applied to the transplant center evaluation data. The fixed effects estimations are similar and most of the variables are very significant. The variable selection procedure in the Chapter 3 suggests sex of the patients is not an important predictor for the patient-level five years post-surgery outcome. While the other predictors, such as cold ischemic time, age at transplantation, BMI and year of transplantation, are important predictors.

The first method finds two clusters among the transplant centers using the sequential testing procedure. The 3 transplant centers, #10, #28 and #287, are detected as outperforming when the False Discovery Rate (FDR) is controlled under 10%. The number of patients are 125, 120 and 114 respectively and in total 329 of the patients die within five years after the surgery. The second method also discovers two clusters among the transplant centers. A cluster with 5 transplant centers, #10, #264, #265, #287 and #290 , stands out the in the dendrogram and the model shows the five transplant centers outperforms the other 290. The 5 transplant centers have 125, 3, 4, 114 and 23 patients respectively and in total 11 of them die with in five years after the surgery. The outperforming transplant centers detected by the two models are different. The FDR approach is more conservative and it only selects a transplant center when the sample size is large and the evidence is strong. In contrast, the clustering method is more sensitive to death rate and the 5 selected transplant centers have the lowest death rate.

4.2 Future Work

4.2.1 Top down instead of bottom up

There are many possible extensions for future research. In the algorithm level, it is interesting to investigate using divisive clustering on the random effects rather than the agglomerative clustering used in the current research. The divisive clustering puts all the random effects in one group with a common mean first. Then it sorts the predicted random effects in an increasing order and searches for a split point that would maximize the resulted likelihood after splitting. The random effects are then divided into subgroups by the split

point. Keeping searching for such split points on the subgroups and dividing the subgroups into even smaller subgroups, the random effects will be successively divided into C subgroups for $C = 1, 2, \ldots$. This divisive algorithm can stop at any reasonable number of clusters before the n transplant centers are divided into n subgroups. Thus, a lot of computation time is saved. Furthermore, the agglomerative clustering result may contain some persistent small cluster that never merge with any other clusters until the end. This kind of cluster is usually fake and it will result in choosing the wrong number of clusters. This kind of small cluster is unlikely to happen in the divisive clustering. However, the major difficulty with divisive clustering in random effects is how to sort the random effects and search for the split point since the random effects are unobserved.

4.2.2 Different penalty functions

The penalty function has an important impact on the clustering result. It determines the merging order of the clusters, and different merging order results in very different clustering trees. For example, the clustering algorithm in Chapter 3 chooses to merge two clusters if

$$|\mu(\mathcal{G}_{c_1}) - \mu(\mathcal{G}_{c_2})| < \lambda_\gamma$$

where $\mu(\mathcal{G}_{c_1})$ and $\mu(\mathcal{G}_{c_2})$ are the means of the clusters $\mathcal{G}_{c_1}$ and $\mathcal{G}_{c_2}$. In contrast, Radchenko and Mukherjee (2017) choose to merge if

$$\frac{|\mu(\mathcal{G}_{c_1}) - \mu(\mathcal{G}_{c_2})|}{|\mathcal{G}_{c_1}| + |\mathcal{G}_{c_2}|} < \lambda_\gamma,$$

where $|\mathcal{G}_{c_1}|$ and $|\mathcal{G}_{c_2}|$ denote the cluster size. Comparing to the rule in our algorithm, Radchenko and Mukherjee (2017) tend to merge the large clusters before small clusters. As a consequence, no clusters can be found in their dendrogram. Motivated from the difference, it is not hard to figure out some other possible rules. For example, we may consider merging two clusters if

$$\frac{|\mathcal{G}_{c_1}||\mathcal{G}_{c_2}||\mu(\mathcal{G}_{c_1}) - \mu(\mathcal{G}_{c_2})|^2}{|\mathcal{G}_{c_1}| + |\mathcal{G}_{c_2}|} < \lambda_\gamma.$$

This rule corresponds to merging the two clusters that will end up with minimum likelihood change and is the same as the Ward's criterion(Ward, 1963). Ward's criterion is very popular

in hierarchical clustering and it is very interesting to investigate its performance in clustering under the framework of GLMM.

4.2.3 Model the fixed effects and random effects at the same time

Besides from extending the clustering method, it is also possible to improve the overall estimations precision of transplant center effects. Some transplant centers with a large amount of patients can be modeled as fixed if their estimation is precise. While the other transplant centers with a small amount of patients can be modeled as random so that they can borrow information from each other. Although the model in Chapter 3 includes both fixed effects model and random effects model as its special cases, it cannot model some transplant centers as fixed effects and some others as random effects at the same time. Clearly, if we assume a transplant center effect to be a random variable with a positive variance, it would be a random effect. On the other hand, if we assume a transplant center effect to be fixed, it is equivalent to assuming it to have zero variance. The model in Chapter 3 assumes common variance for all transplant centers. Thus, the transplant centers are fixed effects only when the tuning parameter λ_γ is 0 and the common variance σ^2 is 0. If we further relax the equal variance assumption, it is possible to estimate some transplant centers as fixed effects and some others as random effects at the same time.

Then the following question is how to perform model selection on such models. The choice between random effects and fixed effects model is usually a result of different philosophy or the purpose of application. When there are both fixed effects and random effects for the transplant centers, we need some practical rules to select the models. Since our goal is to minimizing the overall estimation of transplant center effects, we may select a model that minimizes the mean squared error

$$\sum_{i=1}^{n} E(\widehat{\mu}_i - \mu_{i,0})^2$$

where $\mu_{i,0}$ is the true effect for the transplant center i. Define the bias as $E(\widehat{\mu}_i - \mu_{i,0})$ and the variance as $V(\widehat{\mu}_i)$. The mean squared error can be split into two parts,

$$\sum_{i=1}^{n} E(\widehat{\mu}_i - \mu_{i,0})^2 = \sum_{i=1}^{n} \{E(\widehat{\mu}_i - \mu_{i,0})\}^2 + \sum_{i=1}^{n} V(\widehat{\mu}_i).$$

The first term on the right is the square of the bias and the second term on the right is the variance. The fixed effects model has zero bias but a larger variance while the homogeneous random effects model has a larger bias but a smaller variance. A rough estimation for sum of squares of the biases $\sum_{i=1}^{n}\{E(\widehat{\mu}_i-\mu_{i,0})\}^2$ when there are C clusters is $n\widehat{\sigma}_C^2$. An estimation of the variance $\sum_{i=1}^{n} V(\widehat{\mu}_i)$ is

$$\sum_{c=1}^{C}\left(\sum_{i\in\mathcal{G}_c} N_i\right)\frac{1}{\sum_{i\in\mathcal{G}_c} N_i}\sigma_\epsilon^2 = C\sigma_\epsilon^2$$

where σ_ϵ^2 is the unknown model variance and $\frac{1}{\sum_{i\in\mathcal{G}_c} N_i}\sigma_\epsilon^2$ is the approximate variance for $\widehat{\mu}(\mathcal{G}_c)$. The σ_ϵ^2 may be estimated as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{N_i}{B}\sum_{b=1}^{B}\left(\widehat{\mu}_{b,i}-\frac{1}{B}\sum_{b=1}^{B}\widehat{\mu}_{b,i}\right)^2$$

where $\widehat{\mu}_{b,i}$ is the estimate of μ_i based on bootstrap sample. A new merging rule for clustering can be further defined based on the idea of minimizing the mean squared error. Merging $\mathcal{G}_{c_1}$ and $\mathcal{G}_{c_2}$ would increase the sum of squares of biases by

$$\frac{|\mathcal{G}_{c_1}||\mathcal{G}_{c_2}||\mu_{i_1}-\mu_{i_2}|^2}{|\mathcal{G}_{c_1}|+|\mathcal{G}_{c_2}|}$$

and decrease the sum of variances by σ_ϵ^2. Then merge the two clusters with the smallest

$$\frac{|\mathcal{G}_{c_1}||\mathcal{G}_{c_2}||\mu_{i_1}-\mu_{i_2}|^2}{|\mathcal{G}_{c_1}|+|\mathcal{G}_{c_2}|}$$

until $\frac{|\mathcal{G}_{c_1}||\mathcal{G}_{c_2}||\mu_{i_1}-\mu_{i_2}|^2}{|\mathcal{G}_{c_1}|+|\mathcal{G}_{c_2}|} > \sigma_\epsilon^2$ for all $1 \le c_1 < c_2 \le n$. Ideally, the clustering procedure will stop at the optimal number of clusters.

This idea of minimizing the mean squared error on the estimations of random effects is different from the problem of selecting the number of clusters. In the clustering problem, γ_i (i=1,..., n) are directly observed and σ_ϵ^2 is 0. A model with n clusters is resulted when the mean squared error is minimized. While in GLMM, the selected model depends on the model variance σ_ϵ^2.

4.2.4 Nonparametric density based clustering

Clustering methods other than hierarchical clustering can also be considered. A possible choice is to estimate the second derivative of the density function nonparametrically and

cut the data into subgroups by looking for the valleys in the density function. Clearly, if the true density function $g(\gamma)$ has continuous second derivative, then γ_s will be the lowest point of a valley of $g(\gamma)$ if and only if $g''(\gamma_s) = 0$ and $g''(\gamma) \geq 0$ for γ in a small neighborhood of γ_s. All such points γ_s will split the the true density into several unimodal clusters. When the true density is not available, its second derivative $g''(\gamma)$ can be estimated by a local constant or local linear kernel estimator. The split points can then be found through the estimated second derivative. Since the kernel estimator depends on the bandwidth, our estimated split points will also depend on the bandwidth. If the bandwidth is very small, such split points can be found between any two data points. Then, every data point will be a cluster by its own. On the other hand, if the bandwidth is large and the estimated density function become unimodal, no split points will be found and all the data points will be in one group. That is to say, the problem of selecting the number of clusters becomes equivalent to bandwidth selection in this case. There are already many methods available for bandwidth selection, such as minimizing the Mean Integrated Squared Error.

BIBLIOGRAPHY

Agresti, A., Caffo, B., and Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis*, 47(3):639–653.

Ash, A. S., Fienberg, S. E., Louis, T. A., Normand, S.-L. T., Stukel, T. A., and Utts, J. (2012). Statistical issues in assessing hospital performance. Report, Committee of Presidents of Statistical Societies.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated classification likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719 – 725.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):265–285.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122.

Breslow, N. E. . and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82(1):81–91.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Caffo, B., An, M.-W., and Rohde, C. (2007). Flexible random intercept models for binary outcomes using mixtures of normals. *Computational Statistics & Data Analysis*, 51(11):5220–5235.

Chen, H. and Chen, J. (2003). Tests for homogeneity in normal mixtures in the presence of a structural parameter. *Statistica Sinica*, 13(2):351–365.

Chen, J. (1995). Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, 23(1):221–233.

Chen, J. (2017). Consistency of the MLE under mixture models. *Statistical Science*, 32(1):47–63.

Chen, J. and Khalili, A. (2008). Order selection in finite mixture models with a nonsmooth penalty. *Journal of the American Statistical Association*, 103(484):1674–1683.

Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics*, 37(5A):2523–2542.

Chen, J., Li, P., and Fu, Y. (2012). Inference on the order of a normal mixture. *Journal of the American Statistical Association*, 107(499):1096–1105.

Chen, J., Tan, X., and Zhang, R. (2008). Inference for normal mixtures in mean and variance. *Statistica Sinica*, 18(2):443–465.

Chen, J., Zhang, D., and Davidian, M. (2002). A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics*, 3(3):347–360.

Chi, E. C. and Lange, K. (2015). Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013.

Efron, B. (2004). Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, 99(465):96–104.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fang, Y. and Wang, J. (2012). Selection of the number of clusters via the bootstrap method. *Computational Statistics and Data Analysis*, 56(3):468–477.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Gallant, M. D. and Ronald, A. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika*, 80(3):475–488.

Ghidey, W., Lesaffre, E., and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics*, 60(4):945–953.

Habiger, J., Watts, D., and Anderson, M. (2017). Multiple testing with heterogeneous multinomial distributions. *Biometrics*, 73(2):562–570.

Hartigan, J. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley Conference in Honor of J. Neyman and J. Kiefer, 1985*, pages 807–810.

Hathaway, R. J. (1985). A constrained formulation of maximum likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 13(2):795–800.

He, K., Kalbfleisch, J. D., Li, Y., and Li, Y. (2013). Evaluating hospital readmission rates in dialysis facilities; adjusting for hospital effects. *Lifetime Data Analysis*, 19(4):490–512.

Ho, N. and Nguyen, X. (2016). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6):2726–2755.

Hocking, T. D., Joulin, A., Bach, F., and Vert, J.-P. (2011). Clusterpath: An algorithm for clustering using convex fusion penalties. In *28th international conference on machine learning*, page 1, United States.

Huang, H., Li, Y., and Guan, Y. (2014). Joint modeling and clustering paired generalized longitudinal trajectories with application to cocaine abuse treatment data. *Journal of the American Statistical Association*, 109(508):1412–1424.

Huang, T., Peng, H., and Zhang, K. (2017). Model selection for Gaussian mixture models. *Statistica Sinica*, 27(1):147–169.

Hui, F. K. C., Müller, S., and Welsh, A. H. (2016). Joint selection in mixed models using regularized PQL. *Journal of the American Statistical Association*.

Hui, F. K. C., Warton, D. I., and Foster, S. D. (2014). Order selection in finite mixture models: Complete or observed likelihood information criteria? *Biometrika*, 102(3):724–730.

Ibrahim, J. G. ., Zhu, H., Garcia, R. I. ., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, 67(2):495–503.

Ishwaran, H., James, L. F., and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96(456):1316–1332.

Jiang, J. (2013). The subset argument and consistency of MLE in GLMM: Answer to an open problem and beyond. *The Annals of Statistics*, 41(1):177–195.

Kasahara, H. and Shimotsu, K. (2015). Testing the number of components in normal mixture regression models. *Journal of the American Statistical Association*, 110(512):632–1645.

Keribin, C. . (2000). Consistent estimation of the order of mixture models. *The Indian Journal of Statistics, Series A*, 62(1):49–66.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27(4):887–906.

Komárek, A. and Lesaffre, E. (2008). Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution. *Computational Statistics & Data Analysis*, 52(7):3441–3458.

Krumholz, H., Normand, S.-L. T., Galusha, D., Mattera, J., Rich, A., Wang, Y., and Ward, M. (2006a). Risk-adjustment models for ami and hf: 30-day mortality. Report, Centers for Medicare and Medicaid Services.

Krumholz, H. M., Wang, Y., Mattera, J. A., Wang, Y., Han, L. F., Ingber, M. J., Roman, S., and Normand, S.-L. T. (2006b). An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. *Circulation*, 113:1683–1692.

Lee, Y. and Nelder, J. A. (2004). Conditional and marginal models: Another view. *Statistical Science*, 19(2):219–238.

Li, S., Chen, J., Guo, J., Jing, B.-Y., Tsang, S.-Y., and Xue, H. (2015). Likelihood ratio test for multi-sample mixture model and its application to genetic imprinting. *Journal of the American Statistical Association*, 110(510):867–877.

Li, Y., Cai, X., Glance, L. G., Spector, W. D., and Mukamel, D. B. (2009). National release of the nursing home quality report cards: implications of statistical methodology for risk adjustment. *Health Services Research*, 44(1):79–102.

Liang, F. and Zhang, J. (2008). Estimating the false discovery rate using the stochastic approximation algorithm. *Biometrika*, 95(4):961–977.

Lin, X. and Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91(435):1007–1016.

Lindsten, F., Ohlsson, H., and Ljung, L. (2011). Clustering using sum-of-norms regularization: With application to particle filter output computation. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pages 201–204.

Litière, S., Alonso, A., and Molenberghs, G. (2007). Type I and type II error under random-effects misspecification in generalized linear mixed models. *Biometrics*, 63(4):1038–1044.

Liu, X. and Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *The Annals of Statistics*, 31(3):807–832.

Ma, S. and Huang, J. (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, 112(517):410–423.

McCulloch, C. E. (2003). Generalized linear mixed models. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 7:i–84.

McCulloch, C. E. and Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical Science*, 26(3):388–402.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3):318–324.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359.

Naik, P. a., Shi, P., and Tsai, C.-L. (2007). Extending the Akaike information criterion to mixture regression models. *Journal of the American Statistical Association*, 102(477):244–254.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110.

Podani, J. (1989). New combinatorial clustering methods. *Vegetatio*, 81:61–77.

Radchenko, P. and Mukherjee, G. (2017). Convex clustering using l1 fusion penalty. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for non-identifiable distributions. *The Annals of Statistics*, 9(1):225–228.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710.

Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of American Statistical Association*, 107(497):223–232.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 36(2):111–147.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912.

Sun, W., Reich, B. J., Cai, T. T., Guindani, M., and Schwartzman, A. (2015). False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):59–83.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 58(1):267–288.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika*, 97(4):893–904.

Ward, J. H. (1963). Hierarchical grouping to optimize and objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Woo, M.-J. and Sriram, T. N. (2006). Robust estimation of mixture complexity. *Journal of the American Statistical Association*, 101(476):1475–1486.

Wu, C., Kwon, S., Shen, X., and Pan, W. (2016). A new algorithm and theory for penalized regression-based clustering. *Journal of Machine Learning Research*, 17(1):6479–6503.

Wu, C. F. J. (1983). On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):95–103.

Xu, C. and Chen, J. (2015). A thresholding algorithm for order selection in finite mixture models. *Communications in Statistics - Simulation and Computation*, 44(2):433–453.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

APPENDIX. ADDITIONAL MATERIAL FOR CHAPTER 2

This additional material is for Chapter 2. We provide details of the model fitting algorithm in Section A.1, a simulation procedure in Section A.2 to evaluate the asymptotic distribution in Proposition 2.5, technical assumptions and proof of consistency in Section A.3 and proofs of Propositions 2.1–2.6 in Sections A.4–A.8.

A.1 Model fitting using EM algorithm

We now provide the detailed algorithm to maximize the penalized likelihood in Section 2.2.2.

A.1.1 E-Step with Gauss-Hermite quadrature approximation

At the tth iteration of the algorithm, given the parameter value $\boldsymbol{\theta}^{(t-1)}$ from the previous iteration, we first evaluate the following loss function at the E-step

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t-1)}) = \sum_{i=1}^{n} E\big[\ell_{i,comp}(\theta; \boldsymbol{Y}_i, \boldsymbol{X}_i, \gamma_i, \boldsymbol{L}_i) \mid \boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{\theta}^{(t-1)}\big] + \sum_{c=1}^{C} p_n(\sigma_c^2; \widehat{\sigma}^2_{pilot}) \quad \text{(A.1)}$$

where

$$\begin{aligned} & E\left[\ell_{i,comp}(\theta; \boldsymbol{Y}_i, \boldsymbol{X}_i, \gamma_i, \boldsymbol{L}_i) \mid \boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{\theta}^{(t-1)}\right] \\ &= \sum_{c=1}^{C} \int \log f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma; \boldsymbol{\theta}_y) f(\gamma, L_{ic} = 1 \mid \boldsymbol{X}_i, \boldsymbol{Y}_i; \boldsymbol{\theta}^{(t-1)}) d\gamma \\ &\quad + \sum_{c=1}^{C} \int \log f_c(\gamma \mid \mu_c, \sigma_c) f(\gamma, L_{ic} = 1 \mid \boldsymbol{X}_i, \boldsymbol{Y}_i; \boldsymbol{\theta}^{(t-1)}) d\gamma \\ &\quad + \sum_{c=1}^{C} \log \pi_c \int f(\gamma, L_{ic} = 1 \mid \boldsymbol{X}_i, \boldsymbol{Y}_i; \boldsymbol{\theta}^{(t-1)}) d\gamma, \end{aligned}$$

and $f(\gamma, L_{ic} = 1 \mid \boldsymbol{X}_i, \boldsymbol{Y}_i; \boldsymbol{\theta}^{(t-1)})$ is

$$\frac{\pi_c^{(t-1)} f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma; \boldsymbol{\theta}_y^{(t-1)}) \frac{1}{\sigma_c^{(t-1)}} \phi\left(\frac{\gamma - \mu_c^{(t-1)}}{\sigma_c^{(t-1)}}\right)}{\sum_{c=1}^{C} \pi_c^{(t-1)} \int f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma; \boldsymbol{\theta}_y^{(t-1)}) \frac{1}{\sigma_c^{(t-1)}} \phi\left(\frac{\gamma - \mu_c^{(t-1)}}{\sigma_c^{(t-1)}}\right) d\gamma}.$$

Expectation for a function of a Gaussian random variable can be closely approximated by Gauss-Hermite quadrature:

$$\int h(\gamma)\frac{1}{\sigma}\phi\{(\gamma-\mu)/\sigma\}d\gamma \approx \frac{1}{\sqrt{\pi}}\sum_{m=1}^{M} w_m h(\gamma_m)$$

where $h(\gamma)$ is an integrable real valued function, $\gamma_m = \mu + \sqrt{2}\sigma d_m$, $d_1, \ldots, d_M$ are the Gauss-Hermite abscissas and $w_1, \ldots, w_M$ are the corresponding quadrature weights. We find in our numerical studies that using $M = 100$ quadrature points usually provides a close enough approximation. Denote $\gamma^{(c,m)} = \mu_c^{(t-1)} + \sqrt{2}\sigma_c^{(t-1)} d_m$. The Gauss-Hermite approximation for $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t-1)})$ is

$$\begin{aligned}
&\sum_{i=1}^{n} \frac{\sum_{c=1}^{C}\sum_{m=1}^{M} w_m \pi_c^{(t-1)} \log f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma^{(c,m)}; \boldsymbol{\theta}_y) f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma^{(c,m)}; \boldsymbol{\theta}_y^{(t-1)})}{\sum_{c=1}^{C}\sum_{m=1}^{M} w_m \pi_c^{(t-1)} f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma^{(c,m)}; \boldsymbol{\theta}_y^{(t-1)})} \\
&+\sum_{i=1}^{n} \frac{\sum_{c=1}^{C}\sum_{m=1}^{M} w_m \pi_c^{(t-1)} \left[-\frac{1}{2}\log 2\pi\sigma_c^2 - \frac{1}{2}\frac{(\gamma^{(c,m)}-\mu_c)^2}{\sigma_c^2}\right] f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma^{(c,m)}; \boldsymbol{\theta}_y^{(t-1)})}{\sum_{c=1}^{C}\sum_{m=1}^{M} w_m \pi_c^{(t-1)} f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma^{(c,m)}; \boldsymbol{\theta}_y^{(t-1)})} \\
&+\sum_{i=1}^{n} \frac{\sum_{c=1}^{C}\sum_{m=1}^{M} w_m \pi_c^{(t-1)} \log \pi_c f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma^{(c,m)}; \boldsymbol{\theta}_y^{(t-1)})}{\sum_{c=1}^{C}\sum_{m=1}^{M} w_m \pi_c^{(t-1)} f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma^{(c,m)}; \boldsymbol{\theta}_y^{(t-1)})} + \sum_{c=1}^{C} p_n(\sigma_c^2; \widehat{\sigma}_{pilot}^2) \\
= \quad &\sum_{i=1}^{n}\sum_{c=1}^{C}\sum_{m=1}^{M} \omega_{icm} \left\{ \log f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma^{(c,m)}; \boldsymbol{\theta}_y) - \frac{1}{2}\log 2\pi\sigma_c^2 - \frac{1}{2}\frac{(\gamma^{(c,m)}-\mu_c)^2}{\sigma_c^2} + \log \pi_c \right\} \\
&-a_n \sum_{c=1}^{C} \{\widehat{\sigma}_{pilot}^2/\sigma_c^2 + \log(\sigma_c^2/\widehat{\sigma}_{pilot}^2) - 1\},
\end{aligned}$$

where

$$\omega_{icm} = \frac{w_m \pi_c^{(t-1)} f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma^{(c,m)}; \boldsymbol{\theta}_y^{(t-1)})}{\sum_{c=1}^{C}\sum_{m=1}^{M} w_m \pi_c^{(t-1)} f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma^{(c,m)}; \boldsymbol{\theta}_y^{(t-1)})}. \tag{A.2}$$

A.1.2 M-Step

In the M-step, we maximize $\widehat{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t-1)})$ with respect to $\boldsymbol{\theta}$, and update different components of $\boldsymbol{\theta}$ by

$$\begin{aligned}
\pi_c^{(t)} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} \omega_{icm}, \qquad \mu_c^{(t)} = \frac{\sum_{i=1}^{n}\sum_{m=1}^{M} \gamma^{(c,m)} \omega_{icm}}{\sum_{i=1}^{n}\sum_{m=1}^{M} \omega_{icm}}, \\
(\sigma_c^2)^{(t)} &= \frac{\sum_{i=1}^{n}\sum_{m=1}^{M} (\gamma^{(c,m)} - \mu_c^{(t)})^2 \omega_{icm} + 2a_n \widehat{\sigma}_{pilot}^2}{\sum_{i=1}^{n}\sum_{m=1}^{M} \omega_{icm} + 2a_n},
\end{aligned}$$

and obtain $\boldsymbol{\theta}_y^{(t)}$ by maximizing $\sum_{i=1}^{n}\sum_{c=1}^{C}\sum_{m=1}^{M} \omega_{icm} \log f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma^{(c,m)}; \boldsymbol{\theta}_y)$ using iteratively reweighted least squares.

A.1.3 Stopping rule and random effect prediction

Following Booth and Hobert (1999), we stop the EM algorithm at iteration t if

$$\max_l \frac{|\theta_l^{(t)} - \theta_l^{(t-1)}|}{|\theta_l^{(t-1)}| + 0.001} < 0.005,$$

where θ_l is the lth entry in $\boldsymbol{\theta}$.

At convergence, the weight ω_{icm} can be used to calculate some other quantities of interest, such as the marginal likelihood, the posterior probability of γ_i belonging to the cth component and posterior mean of γ_i . For example, we predict γ_i by its posterior mean

$$\int \gamma f(\gamma \mid \boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{\theta}) d\gamma = \frac{\sum_{c=1}^{C} \pi_c \int \gamma f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma; \boldsymbol{\theta}_y) \phi\{(\gamma - \mu_c)/\sigma_c\}/\sigma_c d\gamma}{\sum_{c=1}^{C} \pi_c \int f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma; \boldsymbol{\theta}_y) \phi\{(\gamma - \mu_c)/\sigma_c\}/\sigma_c d\gamma}.$$

Using the Gauss-Hermite approximation, the posterior mean is approximated by

$$\widehat{\gamma}_i = \sum_{c=1}^{C} \sum_{m=1}^{M} \gamma^{(c,m)} \omega_{icm}, \tag{A.3}$$

where ω_{icm} is defined in (A.2) evaluated at $\widehat{\boldsymbol{\theta}}$.

To obtain reasonable initial values for $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_\gamma$, we first run a generalized linear mixed model assuming γ_i's are i.i.d. normal. We use the estimated fixed effects as initial values for $\boldsymbol{\theta}_y$, fit a Gaussian mixture model on the predicted values $\widehat{\boldsymbol{\gamma}}$ and use the results as the initial values for $\boldsymbol{\theta}_\gamma$.

A.2 Simulation Approach for the Asymptotic Distribution in Proposition 2.5

We use the following procedure to simulate the asymptotic distribution in Proposition 2.5 under the hypothesis $H_0 : C_0 = C$.

Step 0 Fit a C-component latent Gaussian mixture model and obtain the reduced model estimator $\widehat{\boldsymbol{\theta}}_{red}$.

Step 1 Calculate $\tilde{\boldsymbol{s}}_i = (\boldsymbol{s}_{\boldsymbol{\eta},i}^{\mathrm{T}}, \tilde{\boldsymbol{s}}_{\boldsymbol{\lambda},i}^{\mathrm{T}})^{\mathrm{T}}$ with $\tilde{\boldsymbol{s}}_{\lambda,i} = \{(\boldsymbol{s}_{\lambda,i}^{(1)})^{\mathrm{T}}, \ldots (\boldsymbol{s}_{\lambda,i}^{(C)})^{\mathrm{T}}\}^{\mathrm{T}}$, where $\boldsymbol{s}_{\eta,i}$ and $\boldsymbol{s}_{\lambda,i}^{(c)}$, $c = 1, \ldots, C$, are the score functions for the restricted full models defined in (11) evaluated

at $\widehat{\boldsymbol{\theta}}_{red}$. *Let*

$$\tilde{\boldsymbol{I}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\boldsymbol{s}}_i(\tilde{\boldsymbol{s}}_i)^T = \begin{pmatrix} \boldsymbol{I}_\eta & \tilde{\boldsymbol{I}}_{\eta\lambda} \\ \tilde{\boldsymbol{I}}_{\lambda\eta} & \tilde{\boldsymbol{I}}_\lambda \end{pmatrix}.$$

be the sample version of $\tilde{\boldsymbol{\mathcal{I}}} = E\tilde{\boldsymbol{s}}_i\tilde{\boldsymbol{s}}_i^T$, *and calculate* $\tilde{\boldsymbol{I}}_{\lambda|\eta} = \tilde{\boldsymbol{I}}_\lambda - \tilde{\boldsymbol{I}}_{\lambda\eta}\boldsymbol{I}_\eta^{-1}(\tilde{\boldsymbol{I}}_{\lambda\eta})^T$. *To improve numerical stability, we check if* $\tilde{\boldsymbol{I}}$ *is an ill conditioned matrix. If so, set the eigenvalues with small absolute values to be a small positive number.*

Step 2 Generate random a vector

$$\boldsymbol{s} = \left\{(\boldsymbol{s}^{(1)})^{\mathrm{T}}, \ldots, (\boldsymbol{s}^{(C)})^{\mathrm{T}}\right\}^{\mathrm{T}} \sim N(0, \tilde{\boldsymbol{I}}_{\lambda|\eta}).$$

Let $\boldsymbol{I}_{\lambda|\eta}^{(c)}$ *be the sub diagonal matrix of* $\tilde{\boldsymbol{I}}_{\lambda|\eta}$ *corresponding to* $\boldsymbol{s}^{(c)}$. *Then*

$$T_C^* = \max\left\{(\boldsymbol{s}^{(c)})^T(\boldsymbol{I}_{\lambda|\eta}^{(c)})^{-1}\boldsymbol{s}^{(c)}, c = 1, \ldots, C\right\}$$

has the same asymptotic distribution as $T_C(\tau)$ *and* $\widetilde{T}_C$.

Step 3 Repeat Step 2 a large number of times and use the empirical distribution of T_C^* *to approximate the asymptotic distribution of* $\widetilde{T}_C$.

A.3 Assumptions and Consistency of the Estimator in section 2.2.2

A.3.1 Assumptions

For simplicity, assume $N_i = n_0$ for $i = 1, \ldots, n$. Let $(\boldsymbol{X}, \boldsymbol{Y})$ be a generic copy of $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ and have a density

$$f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}) = f(\boldsymbol{x}) \int \left\{\prod_{k=1}^{n_0} f(y_k \mid \boldsymbol{x}_k, \gamma; \boldsymbol{\beta}) g(\gamma \mid \boldsymbol{\theta}_\gamma)\right\} d\gamma \tag{A.4}$$

where $\boldsymbol{y} = (y_1, \ldots, y_{n_0})^{\mathrm{T}}$, $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_0})^{\mathrm{T}}$ and $f(\boldsymbol{x})$ is the joint density of $\boldsymbol{X}$. Define metric

$$\delta(\boldsymbol{\theta}', \boldsymbol{\theta}) = \sum_l |\arctan\theta_l' - \arctan\theta_l|$$

where θ_l is the l-th entry of $\boldsymbol{\theta}$. All convergences in the parameter space are defined with respect to δ.

Assumptions 1–5 below are equivalent to those in Kiefer and Wolfowitz (1956) and Hathaway (1985) for the consistency result. Assumption 6 is a regularity assumption on the penalty function used in Chen et al. (2008) and Kasahara and Shimotsu (2015). Assumption 7 and 8 are additional assumptions for Propositions 2.2 and 2.4 respectively.

Assumption 1 $f(\boldsymbol{x},\boldsymbol{y} \mid \boldsymbol{\theta})$ *is a density (the Radon-Nikodym derivative of a probability measure) with respect to a* σ*-finite measure* μ *on the space of* $(\boldsymbol{x},\boldsymbol{y})$.

Assumption 2 (Continuity Assumption) The definition of $f(\boldsymbol{x},\boldsymbol{y} \mid \boldsymbol{\theta})$ *can be extended to the closure of the parameter space* $\bar{\Theta}_C$ *such that, for any* $\boldsymbol{\theta}^*$ *in* $\bar{\Theta}_C$ *and any Cauchy sequence* $\{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots\} \subset \bar{\Theta}_C$, $f(\boldsymbol{x},\boldsymbol{y} \mid \boldsymbol{\theta}_i) \to f(\boldsymbol{x},\boldsymbol{y} \mid \boldsymbol{\theta}^*)$ *if* $\boldsymbol{\theta}_i \to \boldsymbol{\theta}^*$.

Assumption 3 For any $\boldsymbol{\theta} \in \bar{\Theta}_C$ *and any* $\rho > 0$, $\omega(\boldsymbol{x},\boldsymbol{y} \mid \boldsymbol{\theta},\rho)$ *is a measurable function of* $(\boldsymbol{x},\boldsymbol{y})$, *where*

$$\omega(\boldsymbol{x},\boldsymbol{y} \mid \boldsymbol{\theta},\rho) = \sup f(\boldsymbol{x},\boldsymbol{y} \mid \boldsymbol{\theta}'),$$

the supreme being taken over all $\boldsymbol{\theta}'$ *in* $\bar{\Theta}_C$ *for which* $\delta(\boldsymbol{\theta}',\boldsymbol{\theta}) < \rho$.

Assumption 4 (Identifiability Assumption) Identify $\bar{\Theta}_C$ *as the quotient topological space such that*

$$\mathcal{F} = \left\{ \boldsymbol{\theta} \in \bar{\Theta}_C : \int_{-\infty}^{(\boldsymbol{x}',\boldsymbol{y}')} f(\boldsymbol{x},\boldsymbol{y} \mid \boldsymbol{\theta}) d\mu(\boldsymbol{x},\boldsymbol{y}) = \int_{-\infty}^{(\boldsymbol{x}',\boldsymbol{y}')} f(\boldsymbol{x},\boldsymbol{y}, \mid \boldsymbol{\theta}_0) d\mu(\boldsymbol{x},\boldsymbol{y}) \text{ for any } (\boldsymbol{x}',\boldsymbol{y}') \right\}$$

is identified as a single point.

Assumption 5 For any $\boldsymbol{\theta}'$ *in* $\bar{\Theta}_C$,

$$\lim_{\rho \downarrow 0} E_{\boldsymbol{\theta}} \left[log \frac{\omega(\boldsymbol{x},\boldsymbol{y} \mid \boldsymbol{\theta}',\rho)}{f(\boldsymbol{x},\boldsymbol{y} \mid \boldsymbol{\theta})} \right]^+ < \infty,$$

where $E_{\boldsymbol{\theta}}$ *is the expectation under* $f(\boldsymbol{x},\boldsymbol{y} \mid \boldsymbol{\theta})$. *The notation* $[x]^+$ *equals* x *if* $x > 0$ *and 0 otherwise.*

Assumption 6 The penalty function satisfies, (a) $\sup_{\sigma^2>0} \max\{0, p_n(\sigma^2)\} = o(n)$, $p_n(\sigma^2) = o(n)$ *for any fixed* σ^2*; (b) for any* $\sigma \in (0, 8/(nM)]$, $p_n(\sigma^2) \leq 5\{ln(n)\}^2 \ln(\sigma)$ *for sufficient large* n, *where* $M = \sup_{\boldsymbol{x},\boldsymbol{y}} f(\boldsymbol{y} \mid \boldsymbol{x};\boldsymbol{\theta}_0)$*; (c)* $p_n'(\sigma^2) = o_p(n^{1/4})$ *for any fixed* σ^2.

Assumption 7 When the true number of component is $C_0 = 1$, assume that $\boldsymbol{\mathcal{I}} = E\boldsymbol{I}_n$ is a finite, positive definite matrix, where $\boldsymbol{I}_n$ is defined in (A.6).

Assumption 8 When $\boldsymbol{\theta} \in \Theta_C$, assume that $\boldsymbol{\mathcal{I}}^{(c)}$ defined in (A.12) is positive definite, for $c = 1, \ldots, C$.

Remark 4 The continuity assumption (Assumption 2) is not satisfied by the finite Gaussian mixture model on the boundary of the parameters space, since the likelihood diverges ∞ if any $\sigma_c^2 \to 0$. That is the reason that Hathaway (1985) restricted the estimation in the interior of the parameter space. However, in our problem, the finite Gaussian mixture density $g(\gamma)$ is convoluted with proper density $f(\boldsymbol{y} \mid \boldsymbol{x}, \gamma)$ in (A.4). Since the integral is bounded, unbounded likelihood is no longer a concern and the condition is satisfied even on boundary points of $\bar{\Theta}_C$.

Remark 5 Assumption 4 is a modified version of the identifiability assumption in Kiefer and Wolfowitz (1956). The same assumption is used in Hathaway (1985). The consistency result in Proposition 1 means consistently estimating the mixture density rather than the parameters.

A.3.2 Proof of Proposition 2.1

Using similar arguments as in Chen et al. (2008) one can show, as long as the penalty function satisfies Assumption 6, the maximizer of (2) is restricted in an interior region of the parameter space $\bar{\Theta}(\epsilon) = \{\boldsymbol{\theta} \in \bar{\Theta}; \min_c \sigma_c^2 \geq \epsilon\}$ for some positive constant ϵ. Since the penalty term is of order $o(n)$, which is much smaller than the likelihood function, the maximum penalized likelihood estimator $\widehat{\boldsymbol{\theta}}$ in the restricted parameter space belong to the class of *modified maximum likelihood estimator* in Kiefer and Wolfowitz (1956) and the strong consistency of $\widehat{\boldsymbol{\theta}}$ follows from their theory.

A.4 Proof of Proposition 2.2

Denote for convenience $\zeta_i = \prod_{k=1}^{n_0} f(y_{ik} \mid \boldsymbol{x}_{ik}, \gamma_i; \boldsymbol{\theta}_y)$. After fixing $\pi_1 = \tau$, the log likelihood is

$$l_n(\boldsymbol{\theta}) \quad = \quad \sum_{i=1}^{n} \log \int \zeta_i \{\tau f_1(\gamma \mid \mu_1, \sigma_1) + (1-\tau) f_2(\gamma \mid \mu_2, \sigma_2)\} d\gamma.$$

We adopt the re-parameterization of Kasahara and Shimotsu (2015),

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \sigma_1^2 \\ \sigma_2^2 \end{pmatrix} = \begin{Bmatrix} \nu_\mu + (1-\tau)\lambda_\mu \\ \nu_\mu - \tau\lambda_\mu \\ \nu_\sigma + (1-\tau)(2\lambda_\sigma - \frac{1+\tau}{3}\lambda_\mu^2) \\ \nu_\sigma - \tau(2\lambda_\sigma + \frac{2-\tau}{3}\lambda_\mu^2) \end{Bmatrix}, \tag{A.5}$$

collect all parameters except τ into $\boldsymbol{\psi}(\tau) = (\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\lambda}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\eta} = (\boldsymbol{\theta}_y^{\mathrm{T}}, \nu_\mu, \nu_\sigma)^{\mathrm{T}}$ and $\boldsymbol{\lambda} = (\lambda_\mu, \lambda_\sigma)^{\mathrm{T}}$. Denote $\bar{\Theta}_\psi(\tau)$ as the parameter space of $\boldsymbol{\psi}$ corresponding to $\bar{\Theta}_2(\tau)$. Sometimes we suppress the dependence of $\boldsymbol{\psi}(\tau)$ on τ. Under the null hypothesis $C_0 = 1$, $\lambda_\mu = \lambda_\sigma = 0$ and the true parameter vector is $\boldsymbol{\psi}^* = \{(\boldsymbol{\eta}^*)^{\mathrm{T}}, 0, 0\}^{\mathrm{T}}$.

For any multivariate function $f(\boldsymbol{x})$, denote $\nabla_{\boldsymbol{x}^k} f$ as its k-th derivative, which is a multidimensional array. By similar calculations as in Proposition C and equation (29) in the supplementary appendix of Kasahara and Shimotsu (2015), we can show

$$\begin{aligned}
&\nabla_{\lambda_\mu^k, \boldsymbol{\eta}^\ell} l_n(\boldsymbol{\psi}^*, \tau) = 0, \quad \text{for } k = 1, 2, 3 \text{ and } \ell = 0, 1, 2 \ldots; \\
&\nabla_{\lambda_\mu^k} l_n(\boldsymbol{\psi}^*, \tau) = O_p(n^{1/2}), \quad \text{for } k = 4, 5, 6, 7; \\
&\nabla_{\lambda_\sigma \boldsymbol{\eta}^\ell, \tau} l_n(\boldsymbol{\psi}^*) = 0, \quad \text{for } \ell = 0, 1, 2, \ldots; \\
&\nabla_{\lambda_\sigma^k} l_n(\boldsymbol{\psi}^*, \tau) = O_p(n^{1/2}), \quad \text{for } k = 2, 3; \\
&\nabla_{\lambda_\mu \lambda_\sigma^2} l_n(\boldsymbol{\psi}^*, \tau) = O_p(n^{1/2}); \\
&\nabla_{\lambda_\mu^k \lambda_\sigma} l_n(\boldsymbol{\psi}^*, \tau) = O_p(n^{1/2}), \quad \text{for } k = 1, \ldots, 4.
\end{aligned}$$

Denote $g^*(\gamma) = g(\gamma; \boldsymbol{\psi}^*)$ as the true density of γ under the null hypothesis. Using a ninth order Taylor expansion of l_{pen} around $\boldsymbol{\psi}^*$ as in Kasahara and Shimotsu (2015), we

get the following local quadratic approximation to the penalized likelihood

$$l_{pen}(\boldsymbol{\psi},\tau) - l_{pen}(\boldsymbol{\psi}^*,\tau) = \boldsymbol{t}_n(\boldsymbol{\psi},\tau)^{\mathrm{T}}\boldsymbol{S}_n - \frac{1}{2}\boldsymbol{t}_n(\boldsymbol{\psi},\tau)^{\mathrm{T}}\boldsymbol{I}_n\boldsymbol{t}_n(\boldsymbol{\psi},\tau) + R_n(\boldsymbol{\psi},\tau) + \sum_{c=1}^{2}[p_n\{\sigma_c^2(\boldsymbol{\psi},\tau)\} - p_n\{\sigma_c^2(\boldsymbol{\psi}^*,\tau)\}], \quad \text{(A.6)}$$

where $\boldsymbol{t}_n(\boldsymbol{\psi},\tau) = (\boldsymbol{t}_{\boldsymbol{\eta},n}, \boldsymbol{t}_{\boldsymbol{\lambda},n})^{\mathrm{T}}$, $\boldsymbol{S}_n = \sum_{i=1}^n \boldsymbol{s}_i/\sqrt{n}$, $\boldsymbol{I}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{s}_i\boldsymbol{s}_i^{\mathrm{T}}$, $\boldsymbol{s}_i = (\boldsymbol{s}_{\boldsymbol{\eta},i}^{\mathrm{T}}, \boldsymbol{s}_{\boldsymbol{\lambda},i}^{\mathrm{T}})^{\mathrm{T}}$, $\sigma_c^2(\boldsymbol{\psi},\tau)$ is the variance as a function of $\boldsymbol{\psi}$ defined by the reparameterization in (A.5),

$$\boldsymbol{t}_{\boldsymbol{\eta},n} = \sqrt{n}(\boldsymbol{\eta} - \boldsymbol{\eta}*), \quad \boldsymbol{t}_{\boldsymbol{\lambda},n} = \left\{ \begin{array}{c} 6\sqrt{n}\tau(1-\tau)\lambda_\mu\lambda_\sigma \\ \sqrt{n}\tau(1-\tau)(12\lambda_\sigma^2 - \frac{2}{3}(\tau^2-\tau+1)\lambda_\mu^4) \end{array} \right\},$$

$$\boldsymbol{s}_{\boldsymbol{\eta},i} = \left(\begin{array}{c} \boldsymbol{s}_{\boldsymbol{\theta}_y,i} \\ s_{\nu_\mu,i} \\ s_{\nu_\sigma,i} \end{array} \right) = \left\{ \begin{array}{c} \frac{\int(\partial\zeta_i/\partial\boldsymbol{\theta}_y)g^*}{\int\zeta_i g^*} \\ \frac{\int\zeta_i g^* H_i^{1*}}{\int\zeta_i g^*} \\ \frac{\int\zeta_i g^* H_i^{2*}}{\int\zeta_i g^*} \end{array} \right\}, \quad \boldsymbol{s}_{\boldsymbol{\lambda},i} = \left(\begin{array}{c} \frac{\int\zeta_i g^* H_i^{3*}}{\int\zeta_i g^*} \\ \frac{\int\zeta_i g^* H_i^{4*}}{\int\zeta_i g^*} \end{array} \right),$$

$$R_n(\boldsymbol{\psi},\tau) = [O(\|\boldsymbol{\psi}-\boldsymbol{\psi}^*\|) + o(1)] \times O_p[\{1 + \|\boldsymbol{t}_n(\boldsymbol{\psi},\tau)\|^2\}].$$

Here,

$$H_i^{k*} = H^k(\frac{\gamma_i - \mu_\gamma^*}{\sigma_\gamma^*})/\{k!(\sigma_\gamma^*)^k\}$$

where $H^k(x)$ is the kth order Hermite polynomial, e.g. $H^0(x) = 1$, $H^1(x) = x$, $H^2(x) = x^2 - 1$, $H^3(x) = x^3 - 3x$ and $H^4(x) = x^4 - 6x^2 + 3$.

By consistency of the estimator, we can focus on $\boldsymbol{\psi}$ such that $\|\boldsymbol{\psi} - \boldsymbol{\psi}^*\| = o_p(1)$ and hence $R_n(\boldsymbol{\psi},\tau) = o_p(\|\boldsymbol{t}_n(\boldsymbol{\psi},\tau)\|^2)$. By Assumption 6, $p_n'(\sigma^2) = o_p(n^{1/4})$, and by (A.5)

$$p_n\{\sigma_c^2(\boldsymbol{\psi},\tau)\} - p_n\{\sigma_c^2(\boldsymbol{\psi}^*,\tau)\} = o_p(n^{1/4})(|\lambda_\sigma| + \lambda_\mu^2) = o_p\{\|\boldsymbol{t}_n(\boldsymbol{\psi},\tau)\|\}.$$

Therefore, $l_{pen}(\boldsymbol{\psi},\tau) - l_{pen}(\boldsymbol{\psi}^*,\tau)$ is dominated by the quadratic function defined by the first two terms on the right hand side of (A.6). It is then easy to see $\widehat{\boldsymbol{t}}_n = \boldsymbol{t}_n\{\widehat{\boldsymbol{\psi}}(\tau),\tau\}$ that maximizes $l_{pen}(\boldsymbol{\psi},\tau) - l_{pen}(\boldsymbol{\psi}^*,\tau)$ is

$$\widehat{\boldsymbol{t}}_n = \boldsymbol{I}_n^{-1}\boldsymbol{S}_n + o_p(1). \quad \text{(A.7)}$$

Under Assumption 7, $\boldsymbol{\mathcal{I}} = E\boldsymbol{I}_n$ is a positive definite matrix. By the law of large numbers, $\boldsymbol{I}_n \to \boldsymbol{\mathcal{I}}$ in probability. On the other hand, by the central limit theorem, $\boldsymbol{S}_n \to N(0,\boldsymbol{\mathcal{I}})$ in distribution. Therefore, $\widehat{\boldsymbol{t}}_n \to N(0,\boldsymbol{\mathcal{I}}^{-1})$ in distribution, which also implies

$$\widehat{\boldsymbol{\beta}}_{full}(\tau) - \boldsymbol{\beta}_0 = O_p(n^{-1/2}), \quad \widehat{\lambda}_\mu = O_p(n^{-1/8}), \quad \text{and } \widehat{\lambda}_\sigma = O_p(n^{-1/4}).$$

The convergence rate of $\widehat{\boldsymbol{\theta}}_{\gamma,full}(\tau)$ is determined by those of $\widehat{\lambda}_\mu$ and $\widehat{\lambda}_\sigma$.

A.5 Proof of Proposition 2.3

Following arguments in Section A.4, we have

$$\boldsymbol{S}_n \to N(0, \boldsymbol{\mathcal{I}})$$

in distribution, where $\boldsymbol{\mathcal{I}} = E\boldsymbol{I}_n$. Under the full model, for any $\boldsymbol{\psi}$ such that $\boldsymbol{t}_n = O_p(1)$, using the local quadratic approximation (A.6) we have

$$\begin{aligned} 2\{l_n(\boldsymbol{\psi}, \tau) - l_n(\boldsymbol{\psi}^*, \tau)\} &= 2\boldsymbol{t}_n^{\mathrm{T}}\boldsymbol{S}_n - \boldsymbol{t}_n^{\mathrm{T}}\boldsymbol{I}_n\boldsymbol{t}_n + o_p(1) \\ &= 2\boldsymbol{t}_n^{\mathrm{T}}\boldsymbol{S}_n - \boldsymbol{t}_n^{\mathrm{T}}\boldsymbol{\mathcal{I}}\boldsymbol{t}_n + o_p(1). \end{aligned}$$

Let $\widehat{\boldsymbol{\psi}}_{full}(\tau)$ be maximizer of (A.6) under the full model with 2 components, and it is the reparameterized version of $\widehat{\boldsymbol{\theta}}_{full}(\tau)$. By (A.7), $\boldsymbol{t}_n\{\widehat{\boldsymbol{\psi}}_{full}(\tau)\} = \boldsymbol{\mathcal{I}}^{-1}\boldsymbol{S}_n + o_p(1)$ and hence

$$2[l_n\{\widehat{\boldsymbol{\psi}}_{full}(\tau), \tau\} - l_n(\boldsymbol{\psi}^*, \tau)] = \boldsymbol{S}_n^{\mathrm{T}}\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{S}_n + o_p(1). \tag{A.8}$$

Partition $\boldsymbol{S}_n$ into $\begin{pmatrix} \boldsymbol{S}_{\eta,n} \\ \boldsymbol{S}_{\lambda,n} \end{pmatrix}$ according to the partition of $\boldsymbol{\psi}$. With a similar partition to $\boldsymbol{\mathcal{I}}$, we have

$$\boldsymbol{\mathcal{I}}^{-1} = \begin{pmatrix} \boldsymbol{\mathcal{I}}_\eta & \boldsymbol{\mathcal{I}}_{\eta\lambda} \\ \boldsymbol{\mathcal{I}}_{\lambda\eta} & \boldsymbol{\mathcal{I}}_\lambda \end{pmatrix}^{-1} = \begin{Bmatrix} \boldsymbol{\mathcal{I}}_\eta^{-1} + \boldsymbol{\mathcal{I}}_\eta^{-1}\boldsymbol{\mathcal{I}}_{\eta\lambda}\boldsymbol{\mathcal{I}}_{\lambda|\eta}^{-1}\boldsymbol{\mathcal{I}}_{\lambda\eta}\boldsymbol{\mathcal{I}}_\eta^{-1} & -\boldsymbol{\mathcal{I}}_\eta^{-1}\boldsymbol{\mathcal{I}}_{\eta\lambda}\boldsymbol{\mathcal{I}}_{\lambda|\eta}^{-1} \\ (-\boldsymbol{\mathcal{I}}_\eta^{-1}\boldsymbol{\mathcal{I}}_{\eta\lambda}\boldsymbol{\mathcal{I}}_{\lambda|\eta}^{-1})^{\mathrm{T}} & \boldsymbol{\mathcal{I}}_{\lambda|\eta}^{-1} \end{Bmatrix},$$

where $\boldsymbol{\mathcal{I}}_{\lambda|\eta} = \boldsymbol{\mathcal{I}}_\lambda - \boldsymbol{\mathcal{I}}_{\lambda\eta}\boldsymbol{\mathcal{I}}_\eta^{-1}\boldsymbol{\mathcal{I}}_{\eta\lambda}$. Define

$$\boldsymbol{S}_{\lambda|\eta,n} = \boldsymbol{S}_{\lambda,n} - \boldsymbol{\mathcal{I}}_{\lambda\eta}\boldsymbol{\mathcal{I}}_\eta^{-1}\boldsymbol{S}_{\eta,n},$$

and by simple algebra

$$\boldsymbol{S}_n^{\mathrm{T}}\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{S}_n = \boldsymbol{S}_{\eta,n}^{\mathrm{T}}\boldsymbol{\mathcal{I}}_\eta^{-1}\boldsymbol{S}_{\eta,n} + \boldsymbol{S}_{\lambda|\eta,n}^{\mathrm{T}}\boldsymbol{\mathcal{I}}_{\lambda|\eta}^{-1}\boldsymbol{S}_{\lambda|\eta,n}. \tag{A.9}$$

Under the reduced model, $\boldsymbol{\lambda} = 0$, and hence $\boldsymbol{t}_{\lambda n} = \boldsymbol{S}_{\lambda n} = 0$. Using the same local quadratic approximation, for a parameter vector $\boldsymbol{\psi}_{red}$ in the reduced model,

$$2\{l_n(\boldsymbol{\psi}_{red}, \tau) - l_n(\boldsymbol{\psi}^*, \tau)\} = 2\boldsymbol{t}_{\eta n}^{\mathrm{T}}\boldsymbol{S}_{\eta n} - \boldsymbol{t}_{\eta n}^{\mathrm{T}}\boldsymbol{I}_\eta\boldsymbol{t}_{\eta n} + o_p(1).$$

Let $\widehat{\boldsymbol{\psi}}_{red}$ be the estimator that maximizes the reduced model penalized likelihood, then $\boldsymbol{t}_{\eta n}(\widehat{\boldsymbol{\psi}}_{red}) = \boldsymbol{\mathcal{I}}_{\eta}^{-1}\boldsymbol{S}_{\eta n} + o_p(1)$, and

$$2\{l_n(\widehat{\boldsymbol{\psi}}_{red}, \tau) - l_n(\boldsymbol{\psi}^*, \tau)\} = \boldsymbol{S}_{\eta,n}^{\mathrm{T}}\boldsymbol{\mathcal{I}}_{\eta}^{-1}\boldsymbol{S}_{\eta,n} + o_p(1). \tag{A.10}$$

Combining (A.8), (A.9) and (A.10),

$$T_1(\tau) = 2[l_n\{\widehat{\boldsymbol{\psi}}_{full}(\tau), \tau\} - l_n(\widehat{\boldsymbol{\psi}}_{red}, \tau)] = \boldsymbol{S}_{\lambda|\eta,n}^{\mathrm{T}}\boldsymbol{\mathcal{I}}_{\lambda|\eta}^{-1}\boldsymbol{S}_{\lambda|\eta,n} + o_p(1) \to \chi^2(2) \text{ in distribution.}$$

Because $\boldsymbol{S}_{\lambda|\eta,n}$ and $\boldsymbol{\mathcal{I}}_{\lambda|\eta}$ do not depend on τ,

$$\widetilde{T}_1 = \max_{\tau\in\boldsymbol{\mathcal{T}}} T_1(\tau) = \boldsymbol{S}_{\lambda|\eta,n}^{\mathrm{T}}\boldsymbol{\mathcal{I}}_{\lambda|\eta}^{-1}\boldsymbol{S}_{\lambda|\eta,n} + o_p(1) \to \chi^2(2) \text{ in distribution.}$$

A.6 Proof of Proposition 2.4

Denote $\zeta_i = \prod_{k=1}^{n_0} f(y_{ik} \mid \boldsymbol{x}_{ik}, \gamma_i; \boldsymbol{\theta}_y)$ as in Section A.4. Under the local reparameterization in $\mathcal{N}_{C+1}(c, \tau)$ defined in (2.8) and (2.9) in Section 2.3.3, the log likelihood is

$$l_n(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{n} \log \int \zeta_i g_{c,\tau}(\gamma) d\gamma$$

where

$$\begin{aligned}
g_{c,\tau}(\gamma) \;=\;& (\pi_c + \pi_{c+1})\tau f(\gamma \mid \mu_c, \sigma_c) + (\pi_c + \pi_{c+1})(1-\tau) f(\gamma \mid \mu_{c+1}, \sigma_{c+1}) \\
&+ \sum_{c' \neq c} \pi_{c'} f_{c'}(\gamma \mid \mu_{c'}, \sigma_{c'}) \\
=\;& (\pi_c + \pi_{c+1})\tau f\left\{\gamma \mid \nu_\mu + (1-\tau)\lambda_\mu, \nu_\sigma + (1-\tau)(2\lambda_\sigma - \frac{1+\tau}{3}\lambda_\mu^2)\right\} \\
&+ (\pi_c + \pi_{c+1})(1-\tau) f\left\{\gamma \mid \nu_\mu - \tau\lambda_\mu, \nu_\sigma - \tau(2\lambda_\sigma + \frac{2-\tau}{3}\lambda_\mu^2)\right\} \\
&+ \sum_{c' \neq c} \pi_{c'} f_{c'}(\gamma \mid \mu_{c'}, \sigma_{c'}).
\end{aligned}$$

The score function with respect to $\boldsymbol{\psi}(c, \tau)$ is $\boldsymbol{s}_i^{(c)} = (\boldsymbol{s}_{\boldsymbol{\eta},i}^{\mathrm{T}}, (\boldsymbol{s}_{\boldsymbol{\lambda},i}^{(c)})^{\mathrm{T}})^{\mathrm{T}}$, which is defined in (2.10). Define $\boldsymbol{S}_n^{(c)} = n^{-1/2}\sum_{i=1}^n \boldsymbol{s}_i^{(c)}$, $\boldsymbol{I}_n^{(c)} = n^{-1}\sum_{i=1}^n \boldsymbol{s}_i^{(c)}(\boldsymbol{s}_i^{(c)})^{\mathrm{T}}$ and $\boldsymbol{t}_n\{\boldsymbol{\psi}(c,\tau), \tau\} = (\boldsymbol{t}_{\boldsymbol{\eta},n}, \boldsymbol{t}_{\boldsymbol{\lambda},n})^{\mathrm{T}}$ where

$$\boldsymbol{t}_{\boldsymbol{\eta},n} = \sqrt{n}(\boldsymbol{\eta} - \boldsymbol{\eta}*), \quad \boldsymbol{t}_{\boldsymbol{\lambda},n} = \left\{ \begin{array}{c} 6\sqrt{n}\tau(1-\tau)\lambda_\mu\lambda_\sigma \\ \sqrt{n}\tau(1-\tau)(12\lambda_\sigma^2 - \frac{2}{3}(\tau^2 - \tau + 1)\lambda_\mu^4) \end{array} \right\}.$$

Similar to (A.6), we can derive a local quadratic approximation to the likelihood

$$\begin{aligned} l_n\{\boldsymbol{\psi}(c,\tau),\tau\} - l_n(\boldsymbol{\psi}^*) &= \boldsymbol{t}_n\{\boldsymbol{\psi}(c,\tau),\tau\}^{\mathrm{T}}\boldsymbol{S}_n^{(c)} - \frac{1}{2}\boldsymbol{t}_n\{\boldsymbol{\psi}(c,\tau),\tau\}^{\mathrm{T}}\boldsymbol{I}_n^{(c)}\boldsymbol{t}_n\{\boldsymbol{\psi}(c,\tau),\tau\} \\ &\quad + R_{n,c}\{\boldsymbol{\psi}(c,\tau),\tau\}. \end{aligned} \tag{A.11}$$

where $R_n(\boldsymbol{\psi},\tau) = [O(\|\boldsymbol{\psi}-\boldsymbol{\psi}^*\|) + o(1)] \times O_p[\{1+\|\boldsymbol{t}_n(\boldsymbol{\psi},\tau)\|^2\}]$.

Put $\widehat{\boldsymbol{\psi}}_{full}(c,\tau) = \arg\max_{\boldsymbol{\psi}(c,\tau)\in\Theta_\psi(c,\tau)} l_{pen}\{\boldsymbol{\psi}(c,\tau),\tau\}$ and $\widehat{\boldsymbol{t}}_n = \boldsymbol{t}_n\left\{\widehat{\boldsymbol{\psi}}_{full}(c,\tau),\tau\right\}$. Using similar arguments as in Section A.4, we can show that the penalty function is asymptotically negligible when $\boldsymbol{\psi}(c,\tau)$ is in a consistent neighborhood of $\boldsymbol{\psi}^*$. Define

$$\boldsymbol{\mathcal{I}}^{(c)} = E(\boldsymbol{I}_n^{(c)}) = \mathrm{var}(\boldsymbol{s}_i^{(c)}), \tag{A.12}$$

which is positive definite under Assumption 8. It is then easy to see that

$$\widehat{\boldsymbol{t}}_n = (\boldsymbol{\mathcal{I}}^{(c)})^{-1}\boldsymbol{S}_n^{(c)} + o_p(1) \to \mathrm{Normal}\{0,(\boldsymbol{\mathcal{I}}^{(c)})^{-1}\} \text{ in distribution.} \tag{A.13}$$

By the definition of $\boldsymbol{t}_n\{\boldsymbol{\psi}(c,\tau),\tau\}$, we obtain $\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}^* = O_p(n^{-1/2})$, $\widehat{\lambda}_\mu = O_p(n^{-1/8})$ and $\widehat{\lambda}_\sigma = O_p(n^{-1/4})$. Clearly $\widehat{\mu}_{c,full}(c,\tau)$ and $\widehat{\mu}_{c+1,full}(c,\tau)$ converge to the true parameter $\mu_{c,0}$ at a $O_p(n^{-1/8})$ rate. Since the convergence rates for $\widehat{\sigma}^2_{c,full}(c,\tau)$ and $\widehat{\sigma}^2_{c+1,full}(c,\tau)$ are determined by $\widehat{\lambda}_\mu^2$ and $\widehat{\lambda}_\sigma$, they converge to the true parameter $\sigma^2_{c,0}$ in $O_p(n^{-1/4})$ rate. The rest of the parameters in $\widehat{\boldsymbol{\theta}}_{full}(c,\tau)$ converge in a $O_p(n^{-1/2})$ rate.

A.7 Proof of Proposition 2.5

We first derive the asymptotic properties for $T_C(c,\tau)$. By (A.11) and (A.13),

$$2[l_n\{\widehat{\boldsymbol{\psi}}_{full}(c,\tau),\tau\} - l_n(\boldsymbol{\psi}^*)] = (\boldsymbol{S}_n^{(c)})^{\mathrm{T}}(\boldsymbol{\mathcal{I}}^{(c)})^{-1}\boldsymbol{S}_n^{(c)} + o_p(1),$$

where $\boldsymbol{S}_n^{(c)} \to \mathrm{Normal}(0,\boldsymbol{\mathcal{I}}^{(c)})$ in distribution by the central limit theorem.

The reduced model estimator $\widehat{\boldsymbol{\psi}}_{red}(c,\tau)$ is obtained by minimizing the penalized likelihood while restricting $\lambda_\mu = \lambda_\sigma = 0$. by similar derivations under the full model, we get

$$2[l_n\{\widehat{\boldsymbol{\psi}}_{red}(c,\tau),\tau\} - l_n(\boldsymbol{\psi}^*)] = \boldsymbol{S}_{\eta,n}^{\mathrm{T}}\boldsymbol{\mathcal{I}}_\eta^{-1}\boldsymbol{S}_{\eta,n} + o_p(1),$$

where $\boldsymbol{S}_{\eta,n}$ and $\boldsymbol{\mathcal{I}}_{\eta}$ are sub-vector or sub-matrix of $\boldsymbol{S}_n^{(c)}$ and $\boldsymbol{\mathcal{I}}^{(c)}$ as defined in Proposition 2.5.

Using algebra similar to that in Section A.5, we get

$$\begin{aligned} T_C(c,\tau) &= 2[l_n\{\widehat{\boldsymbol{\psi}}_{full}(c,\tau),\tau\} - l_n\{\widehat{\boldsymbol{\psi}}_{red}(c,\tau),\tau\}] \\ &= (\boldsymbol{S}^{(c)}_{\lambda|\eta,n})^{\mathrm{T}}(\boldsymbol{\mathcal{I}}^{(c)}_{\lambda|\eta})^{-1}\boldsymbol{S}^{(c)}_{\lambda|\eta,n} + o_p(1) \\ &\to \chi^2(2) \text{ in distribution.} \end{aligned}$$

Therefore,

$$T_C(\tau) = \max_c T_C(c,\tau) \to \max\{(\boldsymbol{S}^{(c)}_{\lambda|\eta,n})^{\mathrm{T}}(\boldsymbol{\mathcal{I}}^{(c)}_{\lambda|\eta})^{-1}\boldsymbol{S}^{(c)}_{\lambda|\eta,n}, \quad c=1,\ldots,C\} \text{ in distribution.}$$

Since none of the quantities $(\boldsymbol{S}^{(c)}_{\lambda|\eta,n})^{\mathrm{T}}(\boldsymbol{\mathcal{I}}^{(c)}_{\lambda|\eta})^{-1}\boldsymbol{S}^{(c)}_{\lambda|\eta,n}$ depends on τ, $\widetilde{T}_C$ that maximizes $T_C(\tau)$ over any set $\mathcal{T}$ has the same limiting distribution.

A.8 Proof of Proposition 2.6

The FDR for the described procedure is

$$\begin{aligned} FDR &= E\left\{\frac{\sum_i^n I(\delta_i=1, \sum_{c\in\mathcal{C}_0} L_{ic}=1)}{\sum_i^n I(\delta_i=1)} \mid \sum_i^n I(\delta_i=1)>0\right\} pr\left\{\sum_i^n I(\delta_i=1)>0\right\} \\ &= E\left\{\frac{\sum_i^n \delta_i\left(\sum_{c\in\mathcal{C}_0} L_{ic}\right)}{\sum_i^n \delta_i \vee 1}\right\} \\ &= E\left\{\frac{\sum_i^n \delta_i E\left(\sum_{c\in\mathcal{C}_0} L_{ic}=1 \mid \boldsymbol{X}_i,\boldsymbol{Y}_i\right)}{\sum_i^n \delta_i \vee 1}\right\} \\ &= E\left(\frac{\sum_i^n \delta_i lFDR_i}{\sum_i^n \delta_i \vee 1}\right) \\ &= E\left(\frac{\sum_i^k lFDR_{(i)}}{k}\right) \\ &\le \alpha. \end{aligned}$$